Methods for Analyzing Survival and Binary Data in Complex Surveys

Methods for Analyzing Survival and Binary Data in Complex Surveys

A dissertation presented

by

Kevin Andrew Rader

to

The Department of Biostatistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

May 2014

Professor David P. Harrington **Kevin Andrew Rader**

Methods for Analyzing Survival and Binary Data in Complex Surveys

Abstract

Studies with stratified cluster designs, called complex surveys, have increased in popularity in medical research recently. With the passing of the Affordable Care Act, more information about effectiveness of treatment, cost of treatment, and patient satisfaction may be gleaned from these large complex surveys. We introduce three separate methodological approaches that are useful in complex surveys.

In Chapter 1, we propose a method to create a simulated dataset of clustered survival outcomes with general covariance structure based on a set of covariates. These measurements arise in practice if multiple patients are measured for the same doctor (the cluster) across many doctors. The method proposed in this chapter utilizes the fact that Kendall's Tau is invariant to monotonic transformations in order to create the survival times based on an underlying normal distribution, which the practicing statistician is likely to be more comfortable with. Such a simulated dataset of correlated survival times could be useful to calculate sample size, power, or to measure the characteristics of new proposed methodology.

In Chapter 2, we introduce a method to compare censored survival outcomes in two groups for complex surveys based on linear rank tests. Since the risk sets in a complex survey are not well defined, our proposed method instead utilizes the relationship between the score test of a proportional hazard model and the logrank test to develop the approach in these complex surveys. In order to make this method widely useful, we incorporate propensity scores in order to control for possible confounding effects of other covariates across the two groups.

In Chapter 3, we develop a method to reduce bias in a logistic regression model for binary outcome data in complex surveys. Even in large complex surveys, if the domain is small, a small number of successes or failures may be observed.

When this occurs, standard weighted estimating equations (WEE) may produce biased estimates for the coefficients in the logistic regression model. Based on incorporating an adjustment term in the weighted estimating equation, we are able to reduce the first-order bias of the estimates.

Contents

Acknowledgments

My research and dissertation could not have been possible without the guidance, support, and inspiration of many individuals. I would like to firstly and especially thank my advisors David P. Harrington and Stuart R. Lipsitz. Their dedication, time, expertise, and patience are the main reasons that this thesis was possible. I cannot imagine there are more helpful research advisors anywhere. I would also like to thank Michael Parzen for filling the third chair on my committee and providing invaluable guidance throughout the research and writing process. Not only did Michael play a key role in advising the work itself (along with keeping me focused), but he also introduced me to Stuart and his keen research mind. Words cannot express how truly honored and grateful I am for all three of you.

I would like to thank my family for all of their support along the way. The journey was neither quick nor easy, and my family bore the burden of stress along with me. Mom, Dad, David, Matt, and Allison: your love has meant the world to me. And I hope you understand how happy I am that you have been able to enjoy this work along with me.

To all of my friends, in Boston, Richmond, and Philadelphia, who either had to help me cope with the bad times or got to help me celebrate the good: thank you for being there throughout.

Simulating Clustered Survival Data with Proportional Hazards Margins

Kevin Rader, Stuart Lipsitz, David Harrington, Michael Parzen

Department of Biostatistics
Harvard School of Public Health

1.1 Introduction

In this paper we demonstrate a way to simulate clustered survival data with a general covariance structure. Our simulation approach can be can be used to study the finite sample properties of statistical methods for estimating regression parameters as well to perform sample size and power calculations in applications. Clustered survival data arise often in biomedical studies. For example, in a toxicological study, the mice in the same litter (cluster) are given different harmful chemicals, and the time until death is recorded for each mouse in the litter. In a genetic study, the cluster often is the family, and the outcome is the time from birth until high blood pressure for each member. In cancer clinical trials, patients are often randomized to treatment within institution. In this setting, the institutions can be thought of as clusters and patients as units within a cluster; the outcome for each patient is often the time until the tumor recurs or progresses. Another example of clustering occurs when repeated measures are taken on the same subject. Wei, Lin, and Weissfeld (1989) discuss a clinical trial evaluating the effects of different doses of ribavirin (placebo, low-dose, and high-dose) in preventing HIV-1 virus positivity over time in AIDS patients. In this study, there were 36 patients, and blood samples were obtained from each patient at weeks 4, 8, and 12 from randomization. For a blood sample at each occasion, measures of the level of p24 antigen were taken. A p24 blood concentration of greater than 100 picograms per milliliter was used to indicate the presence of HIV-1 virus. At each of the three occasions, the failure time in this setting is the number of days until the virus was detected in the blood sample. The cluster consists of the failure times for the three blood samples from the same patient. This dataset will be used to illustrate the methods presented in this paper, and the data are given in Table 1.3.

A popular approach for analyzing clustered survival data is the marginal approach, in which the survival time for each member of the cluster has its own 'marginal hazard'. Often this marginal hazard is assumed to be of proportional hazards form. For example, in the AIDS study discussed above, it is of interest

to estimate the treatment effect on the marginal hazard at each measurement point: 4, 8, and 12 weeks from randomization. There is a large literature on fitting marginal regression models for censored survival data; for example, see Wei, Lin, and Weissfeld (1989), Shih and Louis (1995), Huster, Brookmeyer, and Self (1989), Liang, Self, and Chang, (1993), Liang, Self, Bandeen-Roche, and Zeger (1995), Cai and Prentice (1995), Prentice and Hsu (1997), and Segal, Neuhaus and James(1997), Jung and Jeong (2003), and Cai et al. (2007).

To study the finite sample properties of the above approaches, one must be able to simulate multivariate survival data with parametric proportional hazards margins. One could simulate such multivariate survival data from various copulas (Hougaard, 2000). Except for the special case of an exchangeable correlation structure (Marshall and Olkin, 1988; McNeil, Frey and Embrechts, 2005), in which the correlation between all pairs of outcomes within a cluster is the same, we found no papers in the statistical literature on methods to simulate correlated survival data with proportional hazards marginal distributions and general correlation structures. In the Wei, Lin, and Weissfeld (1989) AIDS study discussed above, we would expect the correlation between survival outcomes at 4 and 8 weeks to be larger than the correlation between survival outcomes at 4 and 12 weeks; thus an exchangeable correlation would likely not be appropriate if we wanted to simulate data similar to this AIDS study. Another possibility is the multivariate positive stable distribution (Hougaard, 1986), but again, except for an exchangeable correlation structure (Chambers, Mallows, and Stuck, 1976; Segal, Neuhaus and James, 1997), we found no papers in the statistical literature on methods to simulate multivariate survival data from a positive stable distribution with general correlation structures.

Our approach can also be used to perform sample size and power calculations for studies in which there is expected to be clustered or multivariate survival data. In the current statistical literature, sample size calculations for clustered survival data only reflect an exchangeable correlation structure and has been primarily based on rank tests, for example, see Freedman (1982), Schoenfeld (1983), Manatunga and Chen (2000), Xie and Waksman (2003) and Jung (2007). Our

approach extends these earlier works by basing the simulation on more generalized proportional hazards models and makes no such restrictions on the structure of the correlation or intended analysis.

In this paper, we propose a two-step procedure to simulate correlated survival data. First, we simulate data from a multivariate normal distribution, in which we specify a given correlation matrix, and specify the marginal normal random variables as $N(0,1)$. Next, we use the probability integral transform to transform the marginal normal random variables to the appropriate marginal proportional hazards random variables, which we illustrate with the exponential distribution here. Thus, to perform the simulations, we must specify a correlation matrix for the multivariate normal distribution, as well as the appropriate marginal distribution, here we discuss proportional hazards, but in principle, and parametric (non-proportional hazards) distribution could be specified. Further, as suggested by Hougaard (2000), since Kendall's τ is calculated on rank orders, it should be used as the measure of association between the correlated survival times since it is invariant to transformations of the survival time. Thus, as discussed in Hougaard's book, Kendall's τ between a pair of survival times is a simple linear function of the correlation between the pair of underlying normal random variables. This is very important since, to simulate the correlated survival data, one needs only specify τ and the marginal hazard; thus our procedure generates correlation structures using this measure. Section 2 defines the notation, section 3 illustrates the simulation algorithm with the AIDS example given in Wei, Lin, and Weissfeld (1989), and section 4 extends the procedure in estimating sample size and power for a study.

1.2 Notation and Simulation Method

Suppose there are N independent clusters in the study, with n_i members in the i^{th} cluster. Let T_{ik} denote the failure time for the k^{th} member of cluster i, $i = 1, ..., N$; $k = 1, ..., n_i$. For the i^{th} cluster, we can form an $n_i \times 1$ vector of clustered survival times, $\mathbf{T}_i = [T_{i1}, ..., T_{in_i}]'$. Further, the k^{th} member of cluster i has a $(J \times 1)$

covariate vector $\mathbf{Z}_{ik}$, and we let $\mathbf{Z}_i = [\mathbf{Z}_{i1}, ..., \mathbf{Z}_{in_i}]'$ represent the $n_i \times J$ matrix of covariates for the i^{th} cluster. The density of $\mathbf{T}_i$, from which we would like to simulate, is denoted by $f(t_{i1}, ..., t_{in_i}|\mathbf{Z}_i)$. Further, let $f(t_{ik}|\mathbf{Z}_{ik})$ be the marginal proportional hazards distribution of T_{ik}.

Suppose we let Y_{ik} denote an underlying $N(0,1)$ random variable for the k^{th} member of cluster i. We can form an $n_i \times 1$ vector, $\mathbf{Y}_i = [Y_{i1}, ..., Y_{in_i}]'$, which is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{\Sigma}$, where the diagonal elements of $\mathbf{\Sigma}$ equal 1. Since the diagonal elements of $\mathbf{\Sigma}$ equal 1, $\mathbf{\Sigma}$ is also the correlation matrix of the elements of $\mathbf{Y}_i$. Then, we let

$$\rho_{ijk} = \text{Cov}(Y_{ij}, Y_{ik}) = \text{Corr}(Y_{ij}, Y_{ik}) \tag{1.1}$$

be the correlation between Y_{ij} and Y_{ik} from the multivariate normal distribution. Using the probability integral transform (Hoel, et al., 1971), we know that $U_{ik} = \Phi(Y_{ik})$ has a uniform (0,1) distribution, where $\Phi(\cdot)$ is the cumulative distribution function of the $N(0,1)$ distribution. Applying the probability integral transform once more, $F(T_{ik}|\mathbf{Z}_{ik})$ also has a uniform (0,1) distribution, where $F(t_{ik}|\mathbf{Z}_{ik})$ is the cumulative distribution function of T_{ik}. It then follows that $T_{ik} = F^{-1}(U_{ik}) = F^{-1}(\Phi(Y_{ik}))$ has density $f(t_{ik}|\mathbf{Z}_{ik})$, where $F^{-1}(\cdot|Z_{ik})$ is the inverse cumulative distribution function of T_{ik}. Thus, in summary, $T_{ik} = F^{-1}(\Phi(Y_{ik}))$ will have the proportional hazards distribution of interest, and the T_{ik}'s within the cluster will be correlated since the Y_{ik}'s are correlated. In most computer software packages, one can easily simulate a multivariate normal random vector $\mathbf{Y}_i$. Further, most packages have the standard normal cumulative distribution function built in, as well as most common inverse cumulative distribution functions for proportional hazards models of interest (exponential, Weibull, and extreme value). Further, as discussed by Hougaard (2000), Kendall's τ is recommended as the measure of association between the correlated survival times since it is invariant to transformations of the survival time. As seen in Hougaard (2000), for a pair of normal random variables, Kendall's τ equals

$$\tau_{ijk} = \frac{2\arcsin(\rho_{ijk})}{\pi}, \tag{1.2}$$

where ρ_{ijk} is the correlation between Y_{ij} and Y_{ik} given in (1.1), and $\arcsin(\cdot)$ is the inverse sin function. Since T_{ij} and T_{ik} are just monotone transformations of Y_{ij} and Y_{ik}, and Kendall's τ is invariant to monotone transformations, then (1.2) is also Kendall's τ between T_{ij} and T_{ik}. We suggest first specifying τ_{ijk}, and then transforming back to $\rho_{ijk} = \sin(\pi\tau_{ijk}/2)$ to get the multivariate normal correlation matrix $\mathbf{\Sigma}$. In practice, one may just start by specifying ρ_{ijk} of the multivariate normal distribution.

In summary, to simulate $\mathbf{T}_i$, one can use the following steps:

1. Specify τ_{ijk} and transform to ρ_{ijk} and form $\mathbf{\Sigma}$.
2. Simulate $\mathbf{Y}_i \sim N_{n_i}(\mathbf{0}, \mathbf{\Sigma})$.
3. Calculate $U_{ik} = \Phi(Y_{ik})$.
4. Specify the marginal hazard, and calculate $T_{ik} = F^{-1}(U_{ik}) = F^{-1}(\Phi(Y_{ik}))$.

Finally, the survival time T_{ik} may be right, left, or interval censored, so one can set up the appropriate censoring model of interest, after simulating T_{ik}. In summary, to simulate the correlated survival data, one only really needs to specify Kendall's τ and the marginal hazards.

1.3 A simulation based on the AIDS Clinical Trial

We conducted a simulation experiment based on the ribavirin AIDS clinical trial (Wei, Lin, and Weissfeld, 1989) to explore the computational demands of our approach. The dataset contains $N = 36$ eligible patients (clusters). Each patient was supposed to have a blood sample drawn at weeks 4, 8, and 12 of the trial, although some patients missed visits. Thus, each patient has a maximum of $n_i = 3$ blood samples. The observed response for the k^{th} blood sample from patient i is the minimum of number of days to virus positivity and the censoring time. The main point of interest is the effects of three different doses of ribavirin (placebo, low-dose, and high-dose) on time to virus positivity.

For the k^{th} blood sample from patient i, suppose we assume the time to virus positivity is exponential,

$$f(t_{ik}|\mathbf{Z}_{ik}) = \lambda_{ik} \exp(-\lambda_{ik} t_{ik})$$

with hazard

$$\lambda_{ik}(t) = \lambda_0 \exp(\beta_1 I_{\text{TRT}_1} + \beta_2 I_{\text{TRT}_2} + \beta_3 I_{\text{week}_8} + \beta_4 I_{\text{week}_{12}}), \tag{1.3}$$

where I_{week_8} is the indicator that the observation is from the week 8 observation and $I_{\text{week}_{12}}$ for week 12. Further, in (1.3), I_{TRT_1} is the indicator that the observation was on LOW dose and I_{TRT_2} for HIGH dose of ribavirin. For simplicity in the simulation and without loss of generality we set $\lambda_0 = 1$. Suppose we further assume that Kendall's τ does not depend on individual, but does depend on time, i.e.,

$$(\tau_{i12}, \tau_{i13}, \tau_{i23}) = (\tau_{12}, \tau_{13}, \tau_{23}) = (0.262, 0.128, 0.333).$$

Since $\rho_{ijk} = \sin(\pi \tau_{ijk}/2)$, the correlations for the multivariate normal are:

$$(\rho_{i12}, \rho_{i13}, \rho_{i23}) = (0.4, 0.2, 0.5)$$

When using the simulation method described in the previous section, we have

$$U_{ik} = F(T_{ik}) = 1 - \exp(-\lambda_{ik} T_{ik}),$$

so that

$$T_{ik} = F^{-1}(U_{ik}) = -\lambda_{ik}^{-1} \log(1 - U_{ik}).$$

Table 1.1: Simulation Results based on AIDS Clinical Trial: 1,000 simulated datasets, $N = 300$ patients, and $n_i = 3$. Results calculated using the R programming language. Based on 1000 simulations.

True Parameter Value	$\beta_1 = -0.40$	$\beta_1 = -0.50$	$\beta_2 = 0.10$	$\beta_2 = 0.40$
Mean of parameter estimates	-0.3970	-0.5022	0.1039	0.4060
Mean of variance estimates	0.01114	0.01125	0.00437	0.00576
Observed variance of estimated parameters	0.01127	0.01108	0.00431	0.00543
True coverage of nominal 95% CI	94.7%	94.8%	95.3%	96.0%

Tables 1.4 and 1.5 gives the SAS and R commands, respectively, for a simulation for one patient on each treatment (three patients all together). We note here that we specified the proportional hazards model in (1.3) for simplicity; it does not correspond to the proportional hazards model fitted in the paper by Wei, Lin, and Weissfeld (1989). In the paper by Wei, Lin, and Weissfeld (1989), they fit a Cox proportional hazards model with baseline hazard unspecified (to simulate data, we must specify the baseline hazard).

The results of a larger simulation study are shown in Table 1.1. One thousand datasets were generated with $N = 300$ clusters, and with $n_i = 3$. In particular, we simulated 100 subjects from each of the three dose groups (placebo, LOW dose, and HIGH dose of ribavirin), with $n_i = 3$ repeated survival times on each subject. For the simplicity of this illustration, we chose not to generate any missing data. For each repetition of our simulation, we use the method of Wei, Lin, and Weissfeld (1989), which starts by making a working independence assumption, and fits the usual Cox model to the data, with a robust sandwich variance estimator to consistently estimate the variance. When using the Wei, Lin, and Weissfeld (1989) method, we correctly specified the marginal Cox model as

$$\lambda_{ik}(t) = \lambda_{0k}(t) \exp(\beta_1 I_{\text{TRT}_1} + \beta_2 I_{\text{TRT}_2} + \beta_3 I_{\text{week}_8} + \beta_4 I_{\text{week}_{12}}) \,, \qquad (1.4)$$

When using this marginal Cox model, we will obtain estimates of the treatment effect ($\beta_1 = -0.4$ and $\beta_2 = -0.5$) and the week effects ($\beta_3 = 0.1$ and $\beta_4 = 0.4$), but not the intercept (λ_0). Note, these β_j specifications lead to hazard ratios of approximately 0.670, 0.607, 1.105, and 1.49, for a one unit change in each X_j, respectively, holding the other covariates in the model constant. Wei, Lin, and Weissfeld (1989) showed that if the marginal model is correctly specified, then their estimates of $(\beta_1, \beta_2, \beta_3, \beta_4)$ will be consistent and asymptotically unbiased. The results of the simulation study are given in Table 1.1. In accordance with statistical theory, the Wei, Lin, and Weissfeld (1989) estimates given in Table 1.1 are seen to be unbiased. In fact, the 95% confidence intervals for the week and treatment effects covered the true parameter values 94.7%, 94.8%, 95.3%, and 96.0% of the time. We note that the computational time to simulate all 1000 datasets and calculate the estimates for all 1000 datasets was just under 60 seconds on a 3.2GHz Intel Core i5 (Model 650) computer with 4GB of RAM using R version 3.0.0 for windows. Thus, the proposed procedure to simulate clustered survival data is computationally feasible.

1.4 Use in Sample Size and Power Calculations

This simulation method can also be used to calculate power and sample size prior to a study's inception. Table 1.2 and Figure 1.1 show the results of a power calculation based on a simulation for the AIDS clinical trial data mentioned previously. Again we assume each patient (cluster) has 3 observations. Here the goal was to estimate the power of detecting a treatment effect (β_1 and β_2, which are fixed for a patient) or any week effect (β_3 and β_4, which vary within patients) of based on varying sample sizes (n = 20, 40, 60, 80, and 100 were used). As above we set $\beta_1 = -0.4$, $\beta_2 = -0.5$, $\beta_3 = 0.1$, and $\beta_4 = 0.4$ for this power simulation study. Also, we varied the correlation structure to see how it affected the power calculations, so we chose 4 different sets of specified correlations for the triplets $(\rho_{i12}, \rho_{i13}, \rho_{i23})$ of the original underlying normal distribution: uncorrelated: $(0, 0, 0)$, weak correlation: $(0.2, 0.1, 0.25)$, moderate

Table 1.2: Power calculation for the AIDS Clinical Trial: varying sample size and correlation structure. The tests are 2-parameter Wald tests: (β_1, β_2) for the combined treatment effect and (β_3, β_4) for the combined week effect Results calculated using the R programming language. Based on 1000 simulations.

$(\rho_{i12}, \rho_{i13}, \rho_{i23})$	$(0, 0, 0)$	$(.2, .1, .25)$	$(.4, .2, .5)$	$(.9, .7, .8)$
Treatment Effects				
$N = 20$	0.541	0.453	0.374	0.285
$N = 40$	0.795	0.678	0.605	0.435
$N = 60$	0.908	0.819	0.738	0.566
$N = 80$	0.967	0.905	0.834	0.682
$N = 100$	0.995	0.962	0.911	0.780
Week Effects				
$N = 20$	0.065	0.082	0.162	0.612
$N = 40$	0.110	0.201	0.356	0.888
$N = 60$	0.192	0.314	0.517	0.964
$N = 80$	0.318	0.474	0.684	0.988
$N = 100$	0.440	0.596	0.782	0.993

correlation: $(0.4, 0.2, 0.5)$, and severe correlation: $(0.9, 0.7, 0.8)$. We ran 1000 repetitions of each condition (combination of sample size and underlying correlation structure), and assumed no missing data. The power reported is the proportion of simulations that rejected the null hypothesis at $\alpha = 0.05$ for the robust score test from Wei, Lin, and Weissfeld (1989), which is the default analysis using R's *coxph* function with the *cluster* command.

There are a few points to highlight from the Table 1.2: when there is no intra-cluster correlation (represented by the first column in the table), we see that the power for the two effects are similar (they are different due solely to random variation in the simulation). Not surprisingly, we see that the effect of treatment, β_1 and β_2, which do not vary within patients, has less power to show a true effect as correlation increases. Conversely, the power for the effect of weeks, β_3 and β_4, which do vary within patients, goes up as correlation increases. This follows from previous work (Manatunga and Chen, 2000) that in a cluster-

randomized trial for estimating a treatment effect (a cluster-level effect), for a given size of the treatment effect, when the intra-cluster correlation is high, an investigator needs to have a larger sample size to keep the same power compared to a trial with a lower correlation. For the within cluster effect, this is similar to when the correlation between observations increases in a paired t-test, the power of the test will increase (assuming the difference in means is the same) for each increasing value of correlation. The entire simulation, with a combination of 20 correlation-by-sample size simulations (4 different correlation structures and 5 different sample sizes), took a just over 5 minutes on the same computer mentioned above.

In order to determine an appropriate sample size, one can use a power curve as seen in Figure 1.1 or Figure 1.2. In practice one will need to specify the correlation structure within the clusters (we specified a correlation of $(0.4, 0.2, 0.5)$) along with any possible time effects and the treatment effect (the treatment effects used were $\beta_1 = -0.4$ and $\beta_1 = -0.5$ and the week effects were $\beta_3 = 0.1$ and $\beta_4 = 0.4$). In this example we see that if the desired power was 70% for any effect of treatment in this setting, then the sample size needed was estimated to be $n = 95$ in this case based on Figure 1.1.

As discussed in the Introduction, this approach is very promising since, unlike earlier works, it allows for generalized proportional hazards models and general correlation structures in sample size and power calculations.

Figure 1.1: Power curve for model effects in the AIDS Clinical Trial simulation study varying sample size. Here we combined the treatment effects, $\beta_1 = -0.4$ and $\beta_2 = -0.5$, for one power calculation, and we combined the week effects, $\beta_3 = 0.1$ and $\beta_4 = 0.4$, for a separate calculation. The assumed parameters, (intra-cluster correlations $((\rho_{i12}, \rho_{i13}, \rho_{i23}) = (.4, .2, .5))$, are the same as the simulation in Table 1.1 above. Based on 1000 simulations.

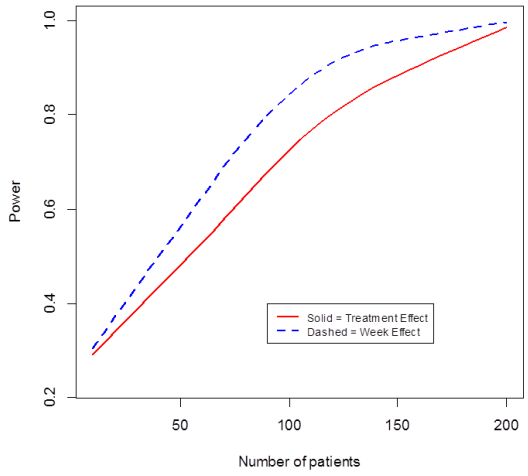
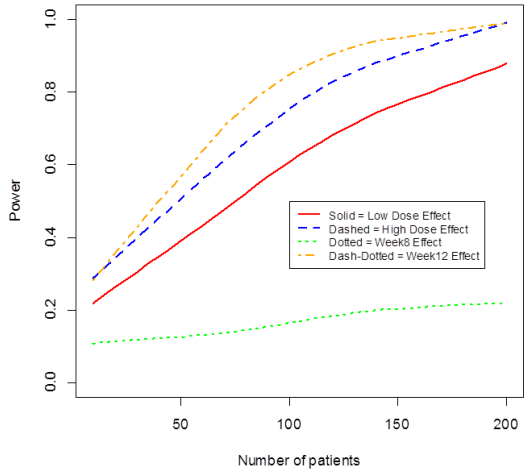


Figure 1.2: Power curve for individual model effects (β_j) in the AIDS Clinical Trial simulation study varying sample size. The assumed parameters are the same as for Figure 1.1 above. Based on 1000 simulations.



1.5 Discussion

We have proposed an approach for simulating correlated survival data with proportional hazards margins. It is simple in that one only needs to specify Kendall's τ and the marginal hazards. An alternative distribution to ours is the positive stable distribution. However, other than exchangeable, the positive stable distribution does not easily generalize to general correlation models. Also, the positive stable only allows positive dependence, but ours allows negative dependence. With some general assumptions on the intra-cluster correlation structure, our method can be used prior to starting a study to determine the power and sample size in these proportional hazards settings. In fact, it is a novel approach when an investigator would like to relax the exchangeable correlation assumption. The approach proposed here is best suited for simulating clustered survival data, but not estimating clustered survival data, as more non-paramteric approaches are available (e.g., Wei, Lin, and Weissfeld, 1989).

Our method is not without weaknesses. In practice it is difficult to get a reasonable correlation to use, unless prior data is available. Most practitioners would prefer to use correlation in terms of ρ, and not Kendall's τ. Our procedure uses a correlation ρ, but it is for the underlying multivariate normal variables, which are not measurable in a real world application, and not the actual survival times. Also, the marginal distribution used in our example, the exponential distribution, has an analytical solution to determine it's inverse cumulative distribution function (CDF). If the inverse CDF had needed to be determined numerically, the procedure would slow down greatly.

Finally we note, the simulation method proposed here can be used to simulate correlated data with any marginal distribution, for example, proportional odds using a log logistic distribution, or very skewed non-proportional hazards distributions like the Pareto. Proportional hazards was used here since that is one of the most common types of analysis used in practice.

Table 1.3: Data from the AIDS Clinical Trial

PATIENT	TREATMENT[a]	WEEK 4	WEEK 8	WEEK 12
1	1	9	6	7
2	1	4	5	10
3	1	6	7	6
4	1	10	-	21∗
5	1	15	8	-
6	1	3	-	6
7	1	4	7	3
8	1	9	12	12
9	1	9	19	19∗
10	1	6	5	6
11	1	9	-	18
12	1	9	20∗	17∗
13	2	6	4	5
14	2	16	17	21∗
15	2	31	19∗	21∗
16	2	27∗	19∗	-
17	2	7	16	23∗
18	2	28∗	7	19∗
19	2	28∗	3	16
20	2	15	12	16
21	2	18	21∗	22
22	2	8	4	7
23	2	4	21∗	7
24	3	21	9	8∗
25	3	13	7	21∗
26	3	16	6	20
27	3	3	8	6
28	3	21	-	25∗
29	3	7	19	3
30	3	11	13	21∗
31	3	27∗	18∗	9
32	3	14	14	6
33	3	8	11	15
34	3	8	4	7
35	3	8	3	9
36	3	19∗	10	17∗

($*$ = censored)
(- = missing)
[a] 1=placebo, 2= low-dose, 3= high-dose

Table 1.4: SAS IML commands for three typical subjects from Table 1.3, for the marginal hazard in (1.3).

```
proc iml;

/* Matrix of Covariates for the 3 subjects, seen 3 times  on the     */
/* three treatments                                                  */
/* column 1 of X is the intercept, col 2 is time, col 3 is treatment */
/* rows 1-3 of X from subject 1,  4-6 from subj 2, 7-9 from subj 3   */

X = { 1 1 0,
      1 2 0,
      1 3 0,
      1 1 1,
      1 2 1,
      1 3 1,
      1 1 2,
      1 2 2,
      1 3 2 };

beta = {1, -1, 1};

lambda = exp(X*beta);

seed = j(9,1,0);     /* random seed (9 x 1) vector of 0's */

Y = RANNOR(seed);    /* 9 independent N(0,1) */

V = {1.0 0.4 0.2,       /* Covariance Matrix for one subject     */
     0.4 1.0 0.5,
     0.2 0.5 1.0};

V = block(V,V,V);      /* Block diagonal covariance for 3 subjects */

call eigen(M,ev,V);
V_1_2 = ev*sqrt(DIAG(M))*ev';    /* Square root matrix of V */

Y = V_1_2*Y ;           /* Y ~ N(0,V)  */

U = CDF('NORMAL',Y) ;

T = -(log(1-U))/lambda;  /* Correlated Exponentials */

print, T;

quit;
```

Table 1.5: R commands for three typical subjects from Table 1.3, for the marginal hazard in (1.3).

```
require(Matrix)
require(mvtnorm)

#########################################################################
# Matrix of Covariates for the 3 subjects: seen 3 times on the 3 trt's #
# column 1 of X is the intercept, col 2 is time, col 3 is treatment    #
# rows 1-3 of X from subject 1, 4-6 from subj 2, 7-9 from subj 3       #
#########################################################################

x1=rep(1,9)
x2=rep(1:3,3)
x3=c(rep(0,3),rep(1,3),rep(2,3))
X=cbind(x1,x2,x3)


beta = c(1,-1,1)
lambda = exp(X%*%beta)


## 9 independent N(0,1)
set.seed(12345)
Y = rnorm(9)


## Covariance Matrix for one subject
v = matrix(c(1.0, 0.4, 0.2, 0.4, 1.0, 0.5, 0.2, 0.5, 1.0),nrow=3)

## Y_i ~ N(0,v)
mvnormals=rmvnorm(nk,sigma=v)
Y = as.vector(t(mvnormals))
U = pnorm(Y)

## Correlated Exponentials
T = -(log(1-U))/lambda
```

Linear Rank Tests for Survival Outcomes in Complex Survey Data

Kevin Rader, Stuart Lipsitz, David Harrington, Michael Parzen

Department of Biostatistics
Harvard School of Public Health

2.1 Introduction

The log rank test and linear rank tests are commonly used statistical tests to determine if there is a difference in survival between two groups. There has been a huge increase in publications on analyses of population-based complex sample surveys in leading medical journals, yet no simple approach has been developed to test for differences between two groups for survival data in this setting. We propose an extension of the linear rank tests for survival outcomes, which is based on the connection between a linear rank test and the score test for the Cox (1975) proportional hazards model. The formulation of our test statistic as a score test statistic from the Cox proportional hazards model for complex survey data paves the way for application of estimating equations score tests, avoiding developing new theory for ranks in complex surveys.

To highlight the use of our method in a real life application, we will be using data from Third National Health and Nutrition Examination Survey (NHANES III), which was conducted by CDC's National Center for Health Statistics.
NHANES III was a multi-stage survey of a representative sample of the US civilian non-institutionalized population. For confidentiality purposes, NHANES gives 49 masked pseudo strata (based on geographic regions) and 98 pseudo primary sampling units, pseudo-PSU's, which can be considered clusters for our purposes. This sampling approach resulted in non-Hispanic blacks, Mexican Americans, and persons over age 60 being over-sampled to obtain reliable information about these subgroups.

We will analyze a subset of this national complex survey, data from $n = 5,532$ hypertensive adults, which was first used by Parikh, 2009. The goal of the original paper was to see if a diet similar to the Dietary Approaches to Stop Hypertension (DASH-like diet) could improve overall survival for hypertensive adults. In the study the Dash-like Diet was ascertained by 24-hour dietary recall using 9 nutrients, and hypertension determined by blood pressure (BP) medication use or measured BP. Overall survival, measured in months, was defined as time from recruitment until death, which was determined from NHANES III Linked Mor-

tality File. Baseline data was collected from 1988 to 1994, with survival being censored in December of 2000.

Several other factors, such as age, sex, race, exercise, or other health habits, were measured that were thought to also affect overall survival. Patient characteristics that were measured at baseline are summarized in Table 3.1, separated by the two treatment-like groups.

In the original paper the primary analysis used was comparing the DASH-like diet vs. standard diet, which we call the *treatment effect*, when controlling for confounding factors listed above. The primary test used was a test to determine whether there was a treatment effect, β_1 , from the following Cox proportional hazard model:

$$\Lambda(t|x_i) = \Lambda_0(t)e^{x_{i,1}\beta_1 + \mathbf{X_{i,2}}\boldsymbol{\beta_2}}$$

where $x_{i,1}$ is an indicator variable for whether they were on the dash diet for the i-th patient and β_1 is the treatment effect, while $\mathbf{X_{i,2}}$ is the vector of possible confounding variables and β_2 is the vector of covariate effects.

Because *a priori* any protective effect of the diet could be accumulated over time, and thus any difference would tend to be seen at the end of the time period, here we suggest weighting treatment differences more towards the end of the time period. In a standard, non-complex survey setting, this would be addressed by using the Harrington and Fleming (1982) class of weighted linear rank tests. As discussed in Klein and Moeschberger (2003), the choice of weights depends on where along the survival function one would like to put the most weight. This paper will outline an extension of this weighted linear rank test for complex surveys.

2.2 General Linear Rank Tests

2.2.1 Survival Data Notation

Consider a typical sampling scheme of n independent subjects $i = 1, 2, ..., n$. We define T_i to be the failure time for the i^{th} individual, and C_i be the failure time

for this individual. We will observe the minimum of T_i and C_i, and hence we can define the censoring indicator for the i^{th} subject, δ_i to be:

$$\delta_i = I\left[T_i \leq C_i\right] = \begin{cases} 1 & \text{if subject } i \text{ is a failure} \\ 0 & \text{if subject } i \text{ is censored} \end{cases}$$

Here our goal is to determine whether the survival time differs between two groups. We define a dichotomous covariate Z_i to be 1 if subject i is in the first group, and 0 if subject i is in the second group. The Cox Model's hazard for subject i at time t given Z_i is then defined to be:

$$\lambda(t|Z_i) = \lambda_0(t)e^{\beta Z_i}$$

where $\lambda_0(t)$ is the arbitrary baseline hazard function. Therefore our main goal of no group difference in hazard rate (and hence survival) would be to test $H_0 : \beta = 0$.

2.2.2 General Score Test

The general form of a score test statistic can be defined as:

$$\chi^2 = \frac{\left[U(\beta = 0)\right]^2}{\left\{\text{Var}\left[U(\beta)\right]\right\}_{\beta=0}}$$

where $U(\beta)$ is the score function (first derivative of the log-likelihood for likelihood based approaches) for β, and $U(\beta = 0)$ is the score function evaluated at $\beta = 0$. $\left\{\text{Var}\left[U(\beta)\right]\right\}_{\beta=0}$ is the variance of $U(\beta)$ is the score function evaluated at $\beta = 0$. For usual maximum likelihood, $\text{Var}\left[U(\beta)\right]$ is estimated by the negative second derivative of the log-likelihood (the observed information). Under the null hypothesis ($\beta = 0$), χ^2 defined above has an asymptotic chi-squared null distribution with one degree of freedom.

Under the assumption of independence between subjects, Cox model's partial likelihood score vector is the the sum over all risk sets, which are the the observed failure times, of the difference between the observed failure Z_i and the average

Z_i in that risk set. This can be written as:

$$U(\beta) = \sum_{i:\delta_i=1} \left\{ Z_i - \bar{Z}_i \right\}$$

where:

$$\bar{Z}_i = \frac{\sum_{j=1}^{n} \left\{ Z_j Y_j(i) e^{\beta Z_j} \right\}}{\sum_{j=1}^{n} \left\{ Y_j(i) e^{\beta Z_j} \right\}}.$$

Thus $\bar{Z}_i$ is a weighted average of the Z_i's in a risk set since

$$Y_j(i) = \begin{cases} 1 & \text{if subject } j \text{ is at risk when subject } i \text{ fails} \\ 0 & \text{otherwise} \end{cases}.$$

For testing the hypothesis $H_0 : \beta = 0$, the numerator of the score test is:

$$\begin{aligned} U(\beta = 0) &= \sum_{i:\delta_i=1} \left\{ Z_i - \bar{Z}_i \right\} \\ &= \sum_{i:\delta_i=1} \left\{ Z_i - \frac{\sum_{j=1}^{n} Z_j Y_j(i)}{\sum_{j=1}^{n} Y_j(i)} \right\} \\ &= \sum_{i:\delta_i=1} \left\{ Z_i - \frac{\sum_{j=1}^{n} Z_j Y_j(i)}{n_i} \right\}, \end{aligned} \tag{2.1}$$

where n_i is the number at risk in the i^{th} risk set. For independent subjects, the score is known to have consistently estimated variance

$$\hat{V}(\beta = 0) = \sum_{i:\delta_i=1} \left\{ \sum_{j=1}^{n} (Z_j - \bar{Z}_i)(Z_j - \bar{Z}_i)' Y_j(i) \right\}.$$

With a little more algebra, this score χ^2 test can be shown to be equivalent to the log rank test.

2.2.3 General Linear Rank Tests

While still assuming independent subjects, for general linear rank tests (Peto and Peto, 1972, test and Harrington and Fleming, 1982, class of G-rho tests) one can use the same Cox score from before, but here the risk sets are *weighted* (Prentice,

1978):

$$U(\beta) = \sum_{i:\delta_i=1} W_i \left\{ Z_i - \bar{Z}_i \right\}$$

W_i is the *analysis* weight to use for the appropriate linear rank test. For example, $W_i = n_i$ for the Wilcoxon test, where n_i is the number of individuals in the i^{th} risk set, $W_i = \hat{S}(t_i)$ for the Peto and Peto test, using $\hat{S}(t_i)$, the Kaplan-Meier (1958) estimated survival collapsed over all groups, and $W_i = \hat{S}(t_{i-1})^p(1 - \hat{S}(t_{i-1}))^q$ for the Harrington and Fleming test.

For independent subjects, the score is known to have consistently estimated variance

$$\hat{V}(\beta = 0) = \sum_{i:\delta_i=1} \left\{ W_i^2 \sum_{j=1}^{n} (Z_j - \bar{Z}_i)(Z_j - \bar{Z}_i)' Y_j(i) \right\}$$

We will use an estimating equations score statistic (Rotnitzky and Jewell, 1990) for testing $H_0 : \beta = 0$ in the model directly above, with some adjustments for complex surveys.

2.3 Extension to Complex Survey Weighting

We let the indicator random variable R_i equal 1 if subject i is selected into the sample and equal 0 otherwise ($i = 1, ..., N$). Thus the probability of being selected into the survey is $P(R_i = 1) = p_i$, which may depend on the outcome of interest, the covariates, or additional variables (screening variables, for example) not in the response model of interest. Each subject in the sample has known weight $w_i = R_i/p_i$.

To adjust for the complex survey sampling, one needs to incorporate this subject-specific sampling weight, w_i, into the linear rank score tests. For complex surveys the linear rank score numerator in this setting generalizes to:

$$U(\beta) = \sum_{i:\delta_i=1} W_i \left\{ w_i Z_i - \bar{Z}_i \right\}$$

where now we define $\bar{Z}_i$ to be:

$$\bar{Z}_i = \sum_{i:\delta_i=1} \left\{ Z_i - \frac{\sum_{j=1}^{n} \{w_j Z_j Y_j(i) e^{\beta Z_j}\}}{\sum_{j=1}^{n} \{w_j Y_j(i) e^{\beta Z_j}\}} \right\}$$

.

The score test statistic is:

$$\chi^2 = \frac{[U(\beta = 0)]^2}{\{\text{Var}\,[U(\beta)]\}_{\beta=0}}$$

Using the results of Binder (1983), the asymptotic variance of $U(\beta = 0)$ is:

$$\{\text{Var}[U(\beta)]\}_{\beta=0} = \left[\text{Var}(\hat{\beta})\right]_{\beta=0} \left[E\left(\frac{dU(\beta)}{d\beta}\right)\right]^2_{\beta=0}, \tag{2.2}$$

where $\left[E\left(\frac{dU(\beta)}{d\beta}\right)\right]_{\beta=0}$ is the negative of the information matrix obtained if one ignores the complex survey design and assumes all subjects are independent with weights w_i. Note, $\text{Var}(\hat{\beta})$ depends on the sample design (stratification, clustering, sampling with or without replacement) as well as the finite population correction factor. Empirically, $\text{Var}(\hat{\beta})$ is estimated via the sandwich variance estimator.

Under the null hypothesis the numerator of the score statistic can be shown to simplify to:

$$U(\beta = 0) = \sum_{i:\delta_i=1} \left\{ W_i \sum_{j=1}^{n} w_j (Z_j - \hat{Z}_j) Y_j(i) \right\}$$

where $\hat{Z}_j = \frac{\sum_{i=1}^{N} Y_j(i) w_i Z_i}{\sum_{i=1}^{N} Y_j(i) w_i}$ is the weighted proportion of subjects in group 1 at risk at time j. Thus, this score statistic for complex survey data can be considered an extension of the usual Linear Rank Test. Most statistical programs for sample surveys allow fitting of proportional hazards models for survival data from complex sample surveys, which makes the implementation more easily widespread. Through asymptotic results and simulation studies we will examine the properties and conclusions of the proposed tests for chosen example complex surveys.

2.4 Incoporating Propensity Scores

One of the biggest drawbacks to using the Linear Rank Tests in observational studies, including complex surveys, is the fact that other covariates can confound the relationship of any group effect on survival. One can use propensity scores to account for possible confounding covariates and to extend the simple linear rank tests to be more widely used in these complex survey applications. There are several ways to incorporate the propensity scores, and there is much debate over what approach is appropriate in any specific setting (Rubin, 1997). Possibly the most straight-forward way is to incorporate yet another weight into the score function of the linear rank test. As shown in Natarajan, et al. (2008) one can just re-weight the score function by the inverse of the estimated propensity scores.

In order to estimate the propensity score, π_i for the i^{th} subject, we fit a logistic regression model to estimate the probability of being assigned to group 1, $Z_i = 1$, for patient i based on a set of covariates, X_i. That is we estimate $\pi_i = \Pr(Z_i = 1|X_i)$ based on a logistic regression model. Note that π_i is the population's theoretical propensity that a patient is assigned to group/treatment Z_i given covariates X_i. We estimate this probability π_i by using weighted logistic regression, using the estimating equations with outcome Z_i and covariates X_i, and using the sampling weights w_i from the complex survey:

$$U(\pi) = \sum_{i=1}^{N} w_i X_i (Z_i - \pi_i)$$

where $\text{logit}(\pi_i) = \beta_0 + \beta X_i$.

In order to incorporate the propensity scores, π_i into the linear rank tests, one needs to multiply the usual score by the inverse of the propensity score, $(1/\pi_i)$ for subjects in the $Z_i = 1$, and multiply by $(1/(1-\pi_i))$ for subjects in the $Z_i = 0$ group. This generalizes to:

$$U(\beta) = \sum_{i:\delta_i=1} W_i \left(w_i^* Z_i - \bar{Z}_i \right)$$

,

where

$$w_i^* = \left(\frac{Z_i}{\pi_i} + \frac{1 - Z_i}{1 - \pi_i} \right) w_i$$

,

and

$$\bar{Z}_i = \sum_{i:\delta_i=1} \left\{ Z_i - \frac{\sum_{j=1}^{n} \left[w_j^* Z_j Y_j(i) e^{\beta Z_j} \right]}{\sum_{j=1}^{n} \left[w_i^* Y_j(i) e^{\beta Z_j} \right]} \right\}$$

Thus the propensity scoreweight can be multiplied with the subject-specific survey weights w_i and be treated as a new subject specific weight in the analysis.

2.5 Application to the DASH-like Diet Study

The goal of the DASH-like diet study from the NHANES complex survey was to compare survival in two groups whose dietary intake differed. The difference in survival between these two pseudo-treatment groups could potentially be confounded with many covariates such as age, sex, race, exercise, or other health habits that could be different in the two pseudo-treatment groups and be related to survival. Any anlysis comparing the diet groups would need to adjust for such covariates.

Applying our methods to the DASH-like Diet requires that first the propensity scores be estimated. The logistic regression propensity model to estimate the log-odds of being assigned to the DASH-like diet group are shown in Table 2.1. Here we see that the DASH-like diet group tended to have healthier lifestyles: activity levels and education level were higher, while rate of smoking and diastolic blood pressure were lower in the DASH-like diet group. Once this logistic model was estimated, then these propensity weights were incorporated into the Cox-based weighted score test as described earlier.

Table 2.1: Coefficients of the Estimated Propensity Model

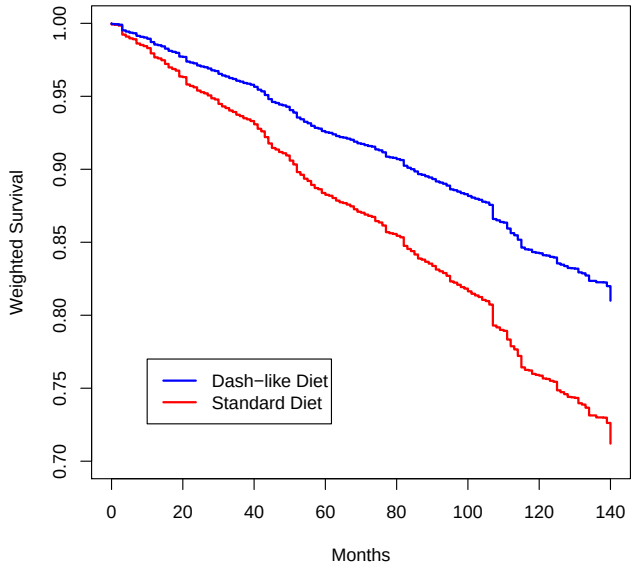
Variable	Estimate	p-value
(Intercept)	-2.425	0.0136
Age	0.0003	0.9400
Sex	-0.1400	0.2189
Race	0.1107	0.0919
Education	0.1769	0.0106
Smoking	-0.5781	0.0004
Obesity	-0.0570	0.7590
Activity	0.3261	<0.0001
CHF	-0.4078	0.0674
BMI	-0.0212	0.1667
MI	0.4437	0.0513
Stroke	0.1384	0.5718
Hyperlipidemia	0.3127	0.0101
BP - Sysolic	0.0008	0.8132
BP - Diastolic	-0.0143	0.0090

The results of various linear rank tests are summarized in Table 2.2. Here we see that the 3 different analysis choices resulted in fairly similar test statistics, which is to be expected based on the weighted survival curves in Figure 2.1, which were estimated based on the methods seen in Xie and Liu, 2005. Each of the three analysis tests ignoring the complex survey design and the correpsonding test taking into accout the study design give quite different results. When the survey design is taken into account, we see more statistically stronger results, whether or not propensity scores are taken into account. And in both cases, the Harrington-Fleming test gives the smallest p-value of a DASH-like diet effect. The results suggest that the effect of the DASH diet may be cumulative over this time frame, as was hypothesized*a priori*.

Table 2.2: Test statistics for various on the DASH-like Diet Study

Type of Linear Rank Test	**Cox Score with No Weighting** χ^2 **(p-value)**	**Survey Weights Only** χ^2 **(p-value)**	**Survey and Propensity Weights** χ^2 **(p-value)**
Logrank	$\chi^2 = 3.10$ $(p = 0.0781)$	$\chi^2 = 8.37$ $(p = 0.0038)$	$\chi^2 = 9.34$ $(p = 0.0022)$
Peto-peto	$\chi^2 = 3.17$ $(p = 0.0752)$	$\chi^2 = 7.60$ $(p = 0.0058)$	$\chi^2 = 8.54$ $(p = 0.0035)$
Harrington-Fleming $p = 0, q = 1$	$\chi^2 = 2.82$ $(p = 0.0933)$	$\chi^2 = 12.36$ $(p = 0.0004)$	$\chi^2 = 13.20$ $(p = 0.0003)$

Figure 2.1: Weighted Kaplan-Meier estimate for all-cause mortality, stratified by diet. The curves are adjusted for confounders based on the propensity scores and are the population estimates incorporating the survey weights.



2.6 Simulation Studies

2.6.1 Proportional Hazard Model

Several simulations were run to confirm the working properties of this linear rank test adaptation to complex surveys; in each a complex survey was simulated. We considered a *true* population survival model to be exponentially distributed with four covariates, giving hazard function:

$$\lambda(t|x_i) = \lambda_0(t)e^{x_{i1}\beta_1+x_{i2}\beta_2+x_{i3}\beta_3+x_{i4}\beta_4}$$

where $x_{i1} \sim Bern(0.3)$ is the strata defining variable in this complex survey setting, x_{i2} is an indicator variable for treatment such that $x_{i2} \sim Bern(0.2)$ when $x_{i1} = 0$ and $x_{i2} \sim Bern(0.7)$ when $x_{i1} = 0$, x_{i3}, and $x_{i4} \sim \chi^2(1)$ represent the potential measured confounding variables for the i-th patient. x_{i3} was defined to vary across these two strata such that for the stratum where $x_{i1} = 0$ we set $x_{i3} \sim N(0, 1)$, and for the stratum where $x_{i1} = 1$ we set $x_{i3} \sim N(1, 1)$. The entire population (N = 1,000,000) was then created based on this model with a specific set of confounding variables in mind. Five different versions of potential confounders were used: no effect of the other covariates: $\beta_1 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$, confounding only by the strata defining variables ($\beta_1 = 0.25$), confounding only by a variable associated to the strata ($\beta_3 = 0.1$), confounding by a variable unrelated to the strata or the treatment (($\beta_4 = 1$), and a combination of all three possible confounding variables ($\beta = (0.25, 0, 0.1, 1)$). The baseline hazard, λ_0, was set to 0.005, and the response variable, y, was right-censored at 144 (to mimic the 12-year follow-up time in the DASH study). No other loss to follow-up was assumed. The observations were then sampled from this population based on a complex scheme: the probability of each observation of being selected, p_{ij} depended on the strata and outcome variable: $p_{ij} = 0.01-0.005(x_1)+0.00001(y)$. This led to an average sample size of 809 observations in the 1000 simulation replications that were used for each simulation setting. The Type I error rates (α level of the tests) for this proportional hazard setting are summarized in Table2.3.

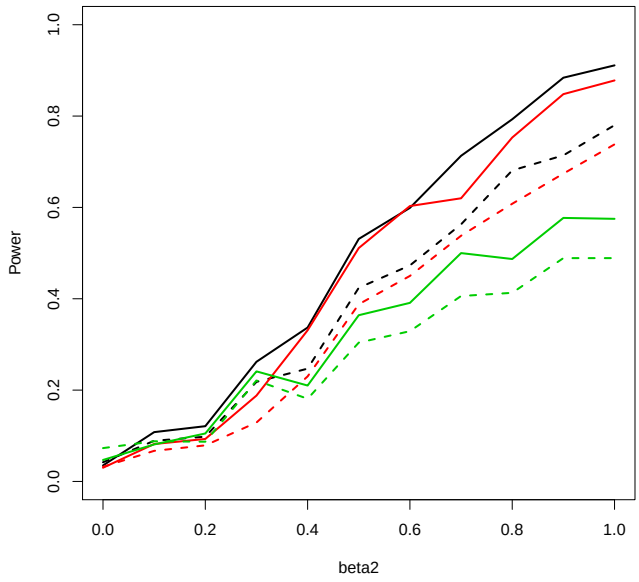
In summary when there are confounding variables present, the tests incorporating the propensity score weights provide approximately correct Type I error rates. This holds for all 3 methods and in both settings of whether the survey weights were used or ignored. Ignoring the propensity score weighting increases the Type I error rates for any test when confounding varaibles are present, especially if those variables are related the probability of being selected into the sample. The inflated Type I error is reduced if the survey weights are used, but it is not eliminated.

A simulation study was also performed to determine the power of rejecting the null hypothesis of no treatment effect, $H_0 : \beta_2 = 0$ for varying levels of β_2 in the six approaches with correct Type I error rates for this proportional hazards model. For this power study we chose the most complicated confounding setting from above: $\beta = (0.25, \beta_2, 0.1, 1)$. The results can be seen in Figure 2.2. In summary, we see that in each case including the survey weights improves the power of each test. Also the Logrank-like test is the most powerful test in this proportional hazards setting.

Table 2.3: Type I error rates for the treatment effect for various tests under specific values of β for the Proportional Hazards simulation. P.A. = Propensity Adjusted and S.W. = Survey Weights used. Based on 1000 replications for each simulation.

Rank Test	**Weighting Used**	$\beta = (0, 0, 0, 0)$	$(0.25, 0, 0, 0)$	$(0, 0, 0.1, 0)$	$(0, 0, 0, 1)$	$(0.25, 0, 0.1, 1)$
Logrank:	**Standard**	0.045	0.217	0.228	0.053	0.509
	P.A.	0.043	0.052	0.051	0.043	0.041
	S.W.	0.047	0.131	0.198	0.054	0.345
	P.A. and S.W.	0.036	0.064	0.052	0.055	0.043
Peto-Peto:	**Standard**	0.050	0.190	0.185	0.047	0.477
	P.A.	0.040	0.056	0.048	0.043	0.040
	S.W.	0.042	0.117	0.172	0.053	0.338
	P.A. and S.W.	0.040	0.056	0.051	0.039	0.035
Harrington-	**Standard**	0.052	0.169	0.192	0.067	0.337
Fleming:	**P.A.**	0.049	0.069	0.058	0.077	0.057
$(p = 0, q = 1)$	**S.W.**	0.054	0.104	0.169	0.061	0.237
	P.A. and S.W.	0.050	0.060	0.059	0.074	0.051

Figure 2.2: Power curve for the Simulated Proportional Hazards Model with varying values of the treatment effect (β_2). The effects of the other confounding variables were set to $\beta = (0.25, \beta_2, 0.1, 0.5)$ with $\lambda_0 = 0.005$. Each line represents a different analysis technique: black lines are for the Logrank-like test, red lines are for the Peto-Peto-like test, and the green lines are for the Harrington-Fleming-like test ($p = 0, q = 1$). The solid lines represent approaches using the survey weights and the dashed lines ignore the survey weights. Based on 1000 replications for each simulation.



2.6.2 Proportional Odds Model

Non-proportional hazardmodels were considered to create survival times for the simulations. These generative models were considered to highlight the importance of using the correct analysis weights. The first non-proprtional hazard model we considered was the proportional odds model. In this form we considered a *true* population survival model with the same four covariates based on a log-logistic form of survival times, which maintains the propotional odds (PO) assumption:

$$\log\left(\frac{1-S(t)}{S(t)}\right) = k\left(\log(\lambda_0 t)\right) + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 + x_{i4}\beta_4$$

where λ and k are the parameters of a log-logistic distribution, respectively. The Wilcoxon or Peto and Peto tests are known to be the optimum linear rank test

under standard survival analysis studies under this proportional odds assumption as these leads to a log-logistic form of survival times (Kalbfleisch and Prentice, 2002; Pettitt, 1984; Bennett, 1983). For this simulation study, the distribution of the covariates, x_{ij}, is identical to the set-up in the PH simulation seen in the previous subsection. Similar to the PH situation above, the entire population (N = 1,000,000) was then created based on this model with a specific set of confounding variables in mind. Five different versions of potential confounders were used: no effect of the other covariates: $\beta_1 = 0$, $\beta_3 = 0$, and $\beta_4 = 0$, confounding only by the strata defining variables ($\beta_1 = 0.25$), confounding only by a variable associated to the strata ($\beta_3 = 0.1$), confounding by a variable unrelated to the strata or the treatment (($\beta_4 = 1$), and a combination of all three possible confounding variables ($\beta = (0.25, 0, 0.1, 1)$). We set $\lambda = 0.01$ and $k = 1.5$ and censored the response variable y at 144. No other loss to follow-up was assumed. The observations were then sampled from this population based on a complex scheme: the probability of each observation of being selected, p_{ij} depended on the strata and outcome variable: $p_{ij} = 0.01 - 0.005(x_1) + 0.00001(y)$. This led to an average sample size of 809 observations in the 1000 simulation replications that were used for each simulation setting. The Type I error rates (α level of the tests) for this proportional odds setting are summarized in Table 2.4. In summary when there are confounding variables present, the tests incorporating the propensity score weights provide approximately correct Type I error rates. This holds for all 3 methods and in both settings of whether the survey weights were used or ignored. Ignoring the propensity score weighting increases the Type I error rates for any test when confounding varaibles are present, especially if those variables are related the probability of being selected into the sample. The inflated Type I error is reduced if the survey weights are used, but it is not eliminated.

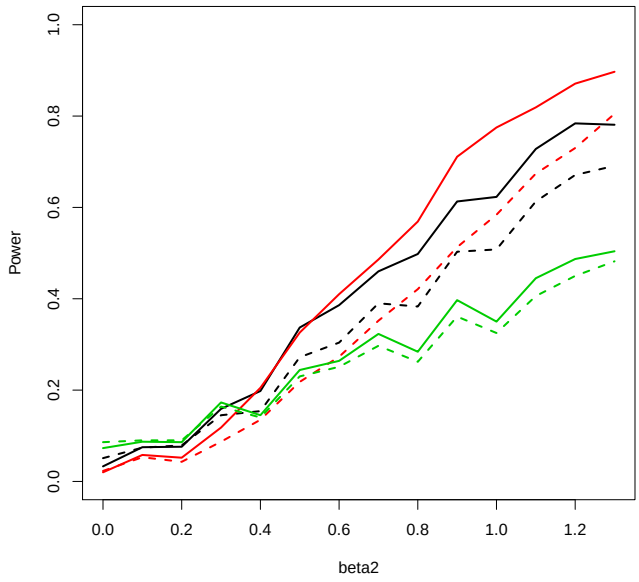
A simulation study was also performed to determine the power of rejecting the null hypothesis of no treatment effect, $H_0 : \beta_2 = 0$ for varying levels of β_2 in the six approaches with correct Type I error rates for this proportional odds model. For this power study we chose the most complicated confounding setting from

above: $\beta = (0.25, \beta_2, 0.1, 1)$. The results can be seen in Figure 2.3. In summary, we see that in each case including the survey weights improves the power of each test. Also the Peto-Peto-like test is the most powerful test in this proportional odds setting.

Table 2.4: Type I error rates for the treatment effect for various tests under specific values of β for the Proportional Odds simulation. P.A. = Propensity Adjusted and S.W. = Survey Weights used. Based on 1000 replications for each simulation.

Rank Test	**Weighting Used**	$\beta = (0,0,0,0)$	$(0.25,0,0,0)$	$(0,0,0.1,0)$	$(0,0,0,1)$	$(0.25,0,0.1,1)$
Logrank:	**Standard**	0.055	0.213	0.476	0.056	0.840
	P.A.	0.047	0.051	0.050	0.040	0.033
	S.W.	0.064	0.106	0.399	0.056	0.699
	P.A. and S.W.	0.059	0.054	0.046	0.046	0.031
Peto-Peto:	**Standard**	0.047	0.249	0.550	0.054	0.893
	P.A.	0.050	0.050	0.043	0.033	0.022
	S.W.	0.056	0.129	0.471	0.052	0.778
	P.A. and S.W.	0.050	0.043	0.041	0.030	0.022
Harrington-	**Standard**	0.059	0.111	0.221	0.054	0.496
Fleming:	**P.A.**	0.057	0.054	0.048	0.063	0.062
$(p = 0, q = 1)$	**S.W.**	0.057	0.066	0.176	0.049	0.363
	P.A. and S.W.	0.054	0.057	0.047	0.061	0.055

Figure 2.3: Power curve for the Simulated Proportional Odds Model with varying values of the treatment effect (β_2). The effects of the other potential confounding variables were set to $\beta = (0.6, \beta_2, 0.3, 1)$ with $\lambda_0 = 0.01$ and $k = 1.5$. Each line represents a different analysis technique: black lines are for the Logrank-like test, red lines are for the Peto and Peto-like test, and the green lines are for the Harrington and Fleming-like test ($p = 0, q = 1$). The solid lines represent approaches using the survey weights and the dashed lines ignore the survey weights. Based on 1000 replications for each simulation.



2.7 Discussion

In summary we have proposed an extension of general linear rank tests for time-to-event data to the complex survey setting. Our approach utilizes the connection between linear rank tests and the score test statistic from the Cox proportional hazard model, and extends this to the complex survey setting which avoids having to develop a new theory for ranks in a complex surveys. An analyst can use our method to compare survival between groups. Our method allows this user to utilize *a priori* hypotheses on how the hazard functions vary over time rather than just use the proportional hazard assumption like in the logrank test. By incorporating the propensity scores into the analysis, one can also adjust for potential measured confounding variables in this setting, which is a novel approach.

Bias Corrected Logistic Regression Models for Complex Surveys

Kevin Rader, Stuart Lipsitz, David Harrington, Michael Parzen

Department of Biostatistics
Harvard School of Public Health

3.1 Introduction

Binary responses are commonplace in studies for many fields: including medical and social sciences. For example, a practicioner may be interested in determining whether or not a patient contracts a disease or complication based on a measureable set of predictors, like age, sex, or environmental exposure factors. The logistic regression model is the most commonly used model for predicting a binary outcome from a set of measurable covariates. Typically, maximum likelihood is the method of choice for estimating the logistic regression model parameters.

However, when the sample size is relatively small or when the binary oucome is either rare or very prevalent, maximum likelihood can yield biased estimates of the logistic regression parameters. In certain cases, when the data has complete or quasi-complete separation, the likelihood may not have a unique solution (Albert and Anderson, 1984). Firth (1993) and Kosmidis and Firth (2009) proposed a procedure to remove the Taylor Series expansion's first-order term in the asymptotic bias of the maximum likelihood estimator. This approach is easily implemented when observations are sampled independently. For the case of logistic regression with independent subjects, there have been numerous methods proposed for handling these data issues, such as exact logistic regression or the bias-correcting approach of Firth (1993); however, such approaches have not been well-studied for binary data from complex sampling schemes. The focus of this paper is on bias-corrected estimates of the regression parameters for the logistic regression model when the data arises from surveys with stratified and clustered designs, often simply referred to as a complex surveys.

Our proposed method is motivated by a study from the 2009 National Inpatient Sample (NIS) that investigated Laparoscopic cystectomies to treat bladder cancer (Yu, et al., 2012), and here we use more recent data (2010) NIS bladder cancer data. Subjects were identified from the US Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample (NIS), sponsored by the Agency for Healthcare Research and Quality (HCUP, 2007). The NIS is a 20% stratified probability sample that encompasses approximately 8 million acute hospital

stays per year from about 1000 hospitals in 45 states. It is the largest all-payer inpatient care observational cohort in the United States and is representative of approximately 90% of all hospital discharges. Based on a similar approach to Yu, et al.(2012), we analyzed patients from the first 6 months of the 2010 NIS that received Laparoscopic cystectomies to treat bladder cancer (n = 385). The primary objective of the study was to compare robot-assisted laparoscopic radical cystectomy (RARC) and open radical cystectomy (ORC) for treatment of bladder cancer. We focus on the primary endpoint of whether or not the patient contracted a wound infection after surgery: $y = 1$ means the patient experienced a wound infection, and $y = 0$ if the patient did not. We want to estimate the difference in the probability of a patient experiencing an infection of the wound area comparing RARC to ORC. There are three a priori potential confounding factors potential associated with wound infection, age, sex, and whether the subject had one or more comorbidities, which are summarized for the two groups in Table 3.1. In our sample from the NIS there were 17 (5.0%) wound complications in the 343 patients who received standard ORC and none of the 42 patients that received robot-assisted treatement, RARC, experienced a wound complication. This leads to the classic issue of separation in the response for these two treatment groups, and motivated us to explore a new analysis approach to handle this issue for the complex survey setting.

In Section 2, we briefly describe the complex sampling design, the typical weighted estimating equations (WEE) for the logistic regression model for complex surveys, and our bias-corrected WEE. In Section 3, we apply this approach for logistic regression analyses of the data from the study of post-operative complications in the laparoscopic cystecomy study (Yu, et al., 2012). In Section 4, we present results of a small-scale simulation study of our bias correction for the logistic regression model. In the example and simulations, we compare our approach to the typical WEE for complex surveys without the bias correction.

Table 3.1: Baseline characteristics bladder cancer patients treated with radical cystectomy in the National Inpatient Sample (NIS).

	Open Radical Cystectomy (ORC), $n = 343$	Robot-assisted Radical Cystectomy (RARC), $n = 42$
Age, years	68.6 (67.6, 69.6)	67.2 (63.3, 71.1)
Female, %	15.2 (12.6, 18.1)	11.9 (5.8, 22.9)
One or more comorbidities, %	22.9 (19.2, 27.0)	21.7 (12.1, 36.0)

Continuous variables are given as means, categorical variables are given as percentages, with ninety-five percent confidence intervals in parentheses. Results are reported as population estimates using survey weights, strata, and cluster variables.

3.2 Methods

3.2.1 Notation for Complex Surveys

The most common type of complex survey design is a stratified cluster design. Further, more complex multi-stage designs can be approximated as a stratified cluster design (Kish, 1965). Thus here we use notation for stratified cluster designs. We let y_{hij} represent the Bernoulli outcome for the j^{th} subject, $(j = 1, \ldots, m_{hi})$, in the i^{th} cluster, $(i = 1, \ldots, n_h)$, within the h^{th} stratum, $(h = 1, \ldots, H)$. Note that we assume there are H strata, n_h clusters in stratum h, and m_{hi} subjects in cluster i of stratum h. Let the indicator variable δ_{hij} equal 1 if subject hij is selected into the sample and equal 0 otherwise. The probability of being selected into the survey is $P(\delta_{hij} = 1) = p_{hij}$ is fixed by the study design and may depend on the outcome of interest, the covariates, or additional variables (screening variables, for example) not in the logistic regression model for the outcome of interest. Thus, each subject in the sample has a known weight $w_{hij} = \delta_{hij}/p_{hij}$. We let π_{hij} be the probability that $Y_{hij} = 1$, which follows the

standard logistic regression model:

$$\pi_{hij} = P(Y_{hij} = 1|\mathbf{x}_{hij}, \beta) = \frac{\exp(\beta'\mathbf{x}_{hij})}{1 + \exp(\beta'\mathbf{x}_{hij})}$$

where $\mathbf{x}_{hij}$ is a $(k+1) \times 1$ vector of covariates including the constant term for the hij^{th} observation, and β is a $(k+1) \times 1$ parameter vector including the intercept term.

To obtain consistent estimates in complex surveys, one needs to incorporate these subject-specific sampling weights, w_{hij}, into the logistic regression estimating equations. Weighting estimating equations (WEE), which naively assume subjects are independent, have been shown to give consistent estimates (Shah, et al., 1996), and are of the form, $U(\hat{\beta}) = 0$, where:

$$U(\beta) = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij} \left(y_{hij} - \pi_{hij}\right) \tag{3.1}$$

Box (1971) showed that typical multivariable estimating equations can be modified to correct for the first-order bias. This can be done by replacing the responses, y_{ihj}, with 'pseudo-response', y^*_{hij}:

$$y^*_{hij} = y_{hij} + a_{hij}$$

where a_{hij} represents the adjustment to the observed response, y_{hij}, and is defined as:

$$a_{hij} = 0.5 \left(\text{tr}\left[\text{Var}(\hat{\beta})D''_{hij}\right]\right) \tag{3.2}$$

where D''_{hij} is the second derivative matrix of the logistic function, π_{hij}, with respect to β.

For the logistic regression model, the contribution of the hij^{th} observation to D''_{hij}

is:

$$\begin{aligned}
D''_{hij} &= \frac{\partial^2}{\partial\beta^2}\left[\pi_{hij}\right] = \frac{\partial^2}{\partial\beta^2}\left(\left[1+\exp(-\mathbf{x}_{hij}\beta)\right]^{-1}\right) \\
&= \frac{\partial}{\partial\beta}\left(\mathbf{x}_{hij}\left[1+\exp(-\mathbf{x}_{hij}\beta)\right]^{-1}\exp(-\mathbf{x}_{hij}\beta)\right) \\
&= \frac{\partial}{\partial\beta}\left(\mathbf{x}_{hij}\pi_i\left[1-\pi_{hij}\right]\right) \\
&= \frac{\partial\pi_{hij}}{\partial\beta}\frac{\partial}{\partial\pi_{hij}}\left(\mathbf{x}_{hij}\pi_i\left[1-\pi_{hij}\right]\right) \\
&= \left(\mathbf{x}_{hij}\pi_{hij}\left[1-\pi_{hij}\right]\right)\left(\left[1-2\pi_{hij}\right]\mathbf{x}'_{hij}\right) \\
&= \mathbf{x}_{hij}\mathbf{x}'_{hij}\pi_{hij}\left(1-\pi_{hij}\right)\left(1-2\pi_{hij}\right),
\end{aligned}$$

and the adjustment factor, a_{hij}, simplifies to:

$$\begin{aligned}
a_{hij} &= 0.5\left(\mathrm{tr}\left[\mathrm{Var}(\hat{\beta})D''_{hij}\right]\right) \\
&= 0.5\left(\mathrm{tr}\left[\mathrm{Var}(\hat{\beta})\mathbf{x}_{hij}\mathbf{x}'_{hij}\pi_{hij}(1-\pi_{hij})(1-2\pi_{hij})\right]\right) \\
&= 0.5\pi_{hij}(1-\pi_{hij})(1-2\pi_{hij})\left[\mathrm{Var}(\mathbf{x}_{hij}\hat{\beta})\right].
\end{aligned}$$

since $\mathrm{Var}(\mathbf{x}_{hij}\hat{\beta})$ is a scalar. Note, in generalized linear model terminology, $\mathrm{logit}(\pi_{hij}) = \mathbf{x}_{hij}\beta$ is the linear predictor. Thus, the adjustment term is a simple function of π_{hij} and the variance of the estimated linear predictor. When there are no sampling weights involved, the adjustment term, a_{hij} is equivalent to Firth's (1993) result for ordinary logistic regression. Replacing y_{hij} in (3.1) with

y^*_{hij}, the bias-reduced estimating equations become:

$$\begin{aligned}
U(\beta)^* &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}\left(y^*_{hij} - \pi_{hij}\right) \\
&= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}\left(y_{hij} + a_{hij} - \pi_{hij}\right) \\
&= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}\left\{y_{hij} + 0.5\pi_{hij}(1-\pi_{hij})(1-2\pi_{hij})\left(\left[\mathrm{Var}(\mathbf{x}_{hij}\hat{\beta})\right]\right) - \pi_{hij}\right\} \\
&= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}\left\{y_{hij} - \pi_{hij} + \left[\pi_{hij}(1-\pi_{hij})\left(\left[\mathrm{Var}(\mathbf{x}_{hij}\hat{\beta})\right]\right)\right](0.5 - \pi_{hij})\right\} \\
&= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}\mathbf{x}_{hij}\left\{y_i - \pi_{hij} + q_{hij}(0.5 - \pi_{hij})\right\}
\end{aligned} \tag{3.3}$$

where $q_{hij} = \pi_{hij}(1-\pi_{hij})\left[\mathrm{Var}(\mathbf{x}_{hij}\hat{\beta})\right]$. For standard logistic regression, the term q_{hij} reduces to the leverage for observation hij, as discussed in Firth (1993) and Heinze and Schemper (2002).

Similar to Heinze and Schemper (2002), bias-corrected estimates can be calculated by splitting each of the original observations into two new observations: one with value y_{hij} and the other with value $1-y_{hij}$ with weights $1+q_{hij}/2$ and $q_{hij}/2$, respectively. Extending these results to complex surveys with weights w_{hij}, we use the weights $w_{hij}(1+q_{hij}/2)$ and $w_{hij}(q_{hij}/2)$ for, y_{hij} and $1-y_{hij}$, respectively. Thus each individual contributes $\{(y_{hij}-\pi_{hij})w_{hij}(1+q_{hij}/2) + (1-y_{hij}-\pi_{hij})w_{hij}(q_{hij}/2)\}$ to the score function, which can be shown to be mathematically

equivalent to the proposed weighted estimating equations:

$$\begin{aligned} U(\beta)^* &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\{(y_{hij}-\pi_{hij})w_{hij}(1+q_{hij}/2)+(1-y_{hij}-\pi_{hij})w_{hij}(q_{hij}/2)\}\mathbf{x}_{hij} \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\{y_{hij}-\pi_{hij}-\pi_{hij}q_{hij}/2+q_{hij}/2--\pi_{hij}q_{hij}/2\}w_{hij}\mathbf{x}_{hij} \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\{q_{hij}-\pi_{hij}-\pi_{hij}q_{hij}+q_{hij}/2\}w_{hij}\mathbf{x}_{hij} \\ &= \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}}\{(y_{hij}-\pi_{hij})+q_{hij}(1/2-\pi_{hij})\}w_{hij}\mathbf{x}_{hij} \end{aligned} \tag{3.4}$$

Even though complex surveys typically have large sample sizes, the issue of separation can occur in large complex surveys when domains are small or subrgoup analyses are performed. The bladder cancer example mentioned in the introduction is an example where this has occurred, as the number of wound infections in the robotic treament arm is zero. In the bias-corrected WEE in (3.4), $y_{hij}=1$ and $y_{hij}=0$ have positive weights, which is equivalent to there being non-zero number of successes ($y_{hij}=1$) and failures ($y_{hij}=0$) at each value of $x_{hij}=1$. Using this property, the results of Wedderburn (1976) can be used to show that the adjusted weighted estimating equations (equivalent to standard logistic regression with weights) we propose has a unique, finite solution (assuming the design matrix is full rank). By splitting each observation into two, we eliminate the problem of separation when the response variable is all successes or all failures for a specific combination of the covariates, and allows for the use of standard complex survey software to do the analysis for example, svyglm in *R*.

The theory by Box (1971) suggests that using a consistent estimate of the true $\text{Var}(\mathbf{x}_{hij}\hat{\beta})$ in q_{hij} should reduce bias.

Two approaches for consistently estimating $\text{Var}(\hat{\beta})$ in q_{hij}, and thus $\text{Var}(\mathbf{x}_{hij}\hat{\beta})$ in q_{hij}, are the typical sandwich estimator and ther small-sample bias-corrected estimator of variance developed by Morel, et al (2003).

To calculate q_{hij}, we consider these two methods of estimating $\text{Var}(\mathbf{x}_{hij}\hat{\beta})$,

along with naive independence. In particular, (a) naively assuming independence among the observations, (b) using the sandwich estimator for $\widehat{\text{Var}}(\mathbf{x}_{hij}\hat{\beta})$ to account for the dependence structure among the observations within a cluster, and (c) a small-sample, bias-corrected sandwich variance estimator proposed by Morel, et al (2003). The robust sandwich estimator of variance used in (b) can be highly variable for rare events or a small number of large clusters, and thus we expect the more stable, bias-corrected corrected sandwhich estimator proposed by Morel, et al (2003) that we use in (c) to lead to less biased estimates than those from the typical sandwich estimator. In fact a priori, we felt using the variance under independence might perform as well as the typical or bias-corrected robust sandwich estimators in this application.

3.2.2 Algorithm for obtaining bias-corrected estimates

To obtain the first-order bias-corrected estimates of β, one can iterate between updating q_{hij} given a current estimate of β and $\widehat{Var}(\mathbf{x}_{hij}\hat{\beta})$, and then re-estimating β and $\widehat{Var}(\mathbf{x}_{hij}\hat{\beta})$ given the updated q_{hij} by solving (3.4), until the estimates of β converge.

In particular, we start by initializing $q_{hij} = k/(\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi})$, which is the average value of the q_{hij} when the observations are independent. Note that $\sum_{h=1}^{H}\sum_{i=1}^{n_h} m_{hi}$ is the total sample size. We then iterate between two steps until convergence of $\hat{\beta}$ is obtained:

1. Calculate the complex survey based estimates of β, but with modified survey weights of $w_{hij}(1+q_{hij}/2)$ for the original y_{hij} and weights of $w_{hij}(q_{hij}/2)$ for the pseudo-observations $1-y_{hij}$ where w_{hij} is the original sampling weight (using svyglm in R or a similar package in another software program).

2. Recalculate q_{hij} based on the estimates, $\hat{\pi}_{hij}$ and $\widehat{Var}(\mathbf{x}_{hij}\hat{\beta})$, from the logistic regression model estimates in the previous step.

Note that in this iterative procedure, the variance estimator used to calculate q_{hij} can be calculated using each of the three proposed approaches mentioned above. However, after convergence, either sandwich variance estimator should be

used to estimate the variance of $\hat{\beta}$ to make inferences. In small samples in which the bias-corrected approach may be warranted, the small-sample, bias-corrected variance estimator of Morel, et al (2003) is the better choice.

3.3 Application to Bladder Cancer Study

In this section, we apply the proposed methods to the analysis of the radical cystectomy data from the National Inpatient Sample (NIS) described in the introduction. This analysis of the NIS includes 385 patients (using the weights, representing 1976 patients in the population) undergoing radical cystectomy to treat bladder cancer throughout the United States. The outcome of interest is binary: whether or not the patient experienced wound infections post-surgery (1=infection, 0 = no infection) while staying at the hospital. Our main comparison of interest is to determine whether the probability of wound infections was different between the two types of cystectomy: standard open radical cystectomy (ORS) and robot-assisted radical cystectomy (RARC). Based on an earlier study (Yu, et al. 2012), a priori, we believed that robot-assisted surgeries would have a lower rate of infection. There are 16 total would infections, all in the ORS group. With 16 total complications, we can have approximatgely 3 covariates based on the results of Vittinghoff and McCulloch (2007). Based on the paper by Yu, et al. (2012), a priori we felt the covariates surgery type (ORC and RARC), age, and sex would be most predictive of wound infection.

To examine the relationship between post-surgery wound infection and these covariates, we fit the logistic regression model,

$$\text{logit}[\pi_i] = \text{logit}\{P[Y_i = 1|\mathbf{x}_i, \beta]\} \quad = \quad \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \ , \tag{3.5}$$

where x_{1i} is the surgery type ($x_{1i} = 1$ for robot-assisted and 0 for standard open radical cystectomy), x_{2i} is the age of the patient, in years, and x_{3i} is the sex of the i^{th} patient ($x_{3i} = 1$ for females and 0 for males).

Table 3.2 gives the estimates of β obtained using the three bias-corrected methods for the data, as well as the standard WEE estimates of β (the latter

were obtained using R `sfyglm`). For comparison, we give the results using all 3 of the approaches to estimate variance used in the calculation of the adjustment factor.

Of note, in Table 3.2, there were no convergence problems with the three bias-corrected methods. However, because there were no complications in the robotic arm, the coefficient for β_1 was converging to $-\infty$; the results based on the WEE model in Table 3.2 are the estimates for the 25^{th} iterations (the default maximum number of iterations in R's `svyglm` function in the package `survey`). Using the independence variance when calculating the adjustment term, we see the estimated odds ratio (OR) to be $e^{-2.774} = 0.062$, controlling for age and sex, when using the robust sandwich estimator of variance the OR is estimated to be $e^{-3.135} = 0.043$, and when the small-sample bias-corrected variance is used the OR is estimated to be $e^{-2.917} = 0.054$. For all four methods, the approach by Morel, et al (2003) was used to calculate all standard error estimates reported in Table 3.2. When comparing the estimates of β to their standard errors, we actually see that the standard WEE produces a much more significant result than the bias-corrected approaches. All 3 of the bias-corrected approaches give very similar estimates, and all lead to the same conclusion if a hypothesis test were conducted at the $\alpha = 0.05$ level. The bias-corrected approach estimating the variance assuming independence or using the Morel, et al (2003) approach to estimate variance is often the most stable, as shown in the next section of this paper.

The other covariates in the model, age, and sex, also give stable estimates of their effects on the probability of wound infection. In this model, the estimated odds ratio is between $e^{-10(0.0330)} = 0.72$ and $e^{-10(0.0193)} = 0.82$ for every 10-year increase in age, and the OR is estimated to be between $e^{0.689} = 1.99$ and $e^{0.808} = 2.24$ for females compared to males. While the bias-corrected approach using the robust sandwich estimate for variance gives significant results for both of these predictors, which is the most unstable of the 3 bias-correected approaches, the other approaches lead to results that are not statistical significant.

In summary, the results of analyses of the bladder cancer data highlight how

the standard WEE approach and the bias-corrected methods can produce discernibly different estimates of effects. However, to examine the finite sample bias of these approaches, we conducted a simulation study; the results of the simulation study are reported in the next section.

Table 3.2: Comparison of WEE logistic regression parameter estimates for the bladder cancer data from the National Inpatient Survey (NIS), $n = 343$.

Effect	Approach	Estimate	SE	Z-statistic	P-value
Intercept	Standard WEE	-1.548	0.868	-1.784	0.079
	Independent Var.	-1.106	1.404	10.817	0.414
Bias-Reduced	Sandwich Var.	-1.436	0.676	-2.237	0.025
	Morel Var.	-1.458	0.909	-1.603	0.109
Robot	Standard WEE	-15.61	0.527	-29.61	<0.001
	Independent Var.	-2.774	1.309	-2.119	0.034
Bias-Reduced	Sandwich Var.	-3.135	0.761	-4.118	<0.001
	Morel Var.	-2.917	1.168	-2.499	0.013
Age	Standard WEE	-0.0325	0.0160	-2.037	0.045
	Independent Var.	-0.0321	0.0191	-1.552	0.121
Bias-Reduced	Sandwich Var.	-0.0330	0.0087	-3.538	0.001
	Morel Var.	-0.0193	0.0177	-1.092	0.278
Female	Standard WEE	0.689	0.430	1.601	0.113
	Independent Var.	0.697	0.564	1.160	0.247
Bias-Reduced	Sandwich Var.	0.703	0.258	2.571	0.010
	Morel Var.	0.808	0.925	0.874	0.385

3.4 Simulation Study

In this section, we study the empirical relative bias in estimating β using typical logistic regression models incorporating the complex survey structure (WEE) and our bias-reduced approach using 3 different variance estimators when calculating the multiplicative weighting factor, q_{hij}: variance under independence (Independence), variance using the robust sandwich estimator (Sandwich), and using Morel, et al. (2003) small-sample variance correction (Morel). For simplicity, in the simulation study, we used a cluster design without stratification and weighting, where sampling of clusters was performed without replacement from a finite population of clusters.

For the simulations, the true marginal logistic model for any subject in the population is

$$\text{logit}(P[Y_{ij} = 1|\mathbf{x}_{ij}]) = \beta_0 + \sum_{k=1}^{10} \beta_k x_{ijk} \ , \tag{3.6}$$

where the ten x_{ijk}'s are independent $Bern(p_x)$ variables. The intercept β_0 was chosen so that the average $P[Y_{ij} = 1]$ equals 0.20. This marginal model is similar to that used in a simulation study performed by Heinze and Schemper (2002). For simplicity, we set all ten β_k equal to the same value.

To simulate the clustered data, we use the random intercept logistic regression model proposed by Wang and Louis (2003) and further developed by Parzen, et al. (2011). In particular, the conditional subject-specific logistic regression model is

$$\text{logit}(P[Y_{ij} = 1|\mathbf{x}_{ij}]) = b_i + \left(\beta_0 + \sum_{k=1}^{10} \beta_k x_{ijk}\right)/\phi, \tag{3.7}$$

where, given the subject-specific random effect b_i, the Y_{ij}'s from the same cluster are independent Bernoulli random variables. When b_i follows a 'bridge' distribution, the marginal logistic regression (Wang and Louis, 2003) equals that given in (3.6). The bridge random variable has mean 0 and ϕ is the rescaling parameter. In particular,

$$\text{Var}(b_i) = \frac{\pi^2}{3}\left(\frac{1}{\phi^2} - 1\right),$$

so that the larger the value of ϕ, $(0 < \phi < 1)$, the smaller the variance (and the lower the correlation between pairs of random variables in the same cluster).

We denote the population number of clusters by N, which we set to $N = 400$, the number of sampled clusters by n, and the cluster size by m_i (we assume all clusters have the same size, and all members of the cluster are sampled).

We conducted 24 simulation configurationss varying the following conditions: the effect of the covariates, $\beta_k = \beta = \{\ln(2), \ln(4), \ln(16)\}$ (recall, we set all ten β_k to the same value); cluster sizes, $m_i = \{5, 10\}$; the bridge distribution's scaling parameter, $\phi = \{0.7, 0.9\}$; and the number of clusters sampled, $n = 40$ and $n = 80$. For each simulation configuration, 2000 simulation replications were performed. The convergence criterion for WEE is that the relative change in the log-likelihood between successive iterations is less than 0.000001; we report the percentage of simulation replications in which this convergence criterion was not met. When the usual WEE fails to converge, we use the estimates from the 25^{th} iteration (the default maximum number of iterations in R's `svyglm` function in the package `survey`).

Tables 3.3, 3.4, and 3.5 present the relative biases for β_2 defined as $100(\hat{\beta}_2 - \beta_2)/\beta_2$, the mean square error of the estimates, and the empirical coverage probabilities of 95% Wald confidence intervals for all the simulation study specifications, respectively. Without loss of generality we report results for β_2; any of the β_k could have been selected for bias reporting since this model is symmetric across covariates (all covariates are independent and have the same Bernoulli distribution and all ten $\beta_k = \beta$), but β_2 was selected to match the approach by Heinze and Schemper (2002). We see here that the relative bias is greatly reduced, by an order of magnitude, when using any of the three bias-reduced approaches in comparison to standard GEE. In Table 3.5 where the effect size, β, is largest and the sample size is small, $m_i = 5$ and $n = 40$, it appears that the robust sandwich estimator does the worst job of the 3 bias-reduced approaches as it has the highest amount of relative bias, mean square error, and coverage probabilties below the nominal 95% level. Based on these results, it is our suggestion that an applied statistician should choose the bias-reduced approach using either the

independent variance estimate or small-sample bias-reduced version of the robust variance estimate to calculate q_{hij} when performing the analysis. The standard WEE approach gave average estimated values for β close to zero, showing no effect when there trule was an effect of at least $\beta = 0.69$. Because of this fact, the average relative bias for the standard WEE method are all very close to -100% in all simulation configurations.

Although Wald confidence intervals are known to be conservative (Hauck and Donner, 1977; Heinze and Schemper, 2002; and Bull, et al., 2007) with large β's, we found in nearly all sets of simulations with $\beta_2 = 2.77$, that the coverage probabilities agree with the nominal 95% level as long as the sample size is large. However, we should not generalize based on this one simulation setup, so one would still want alternatives to obtain confidence intervals.

3.5 Discussion

In this paper we have described a simple implementation of bias correction in the logistic regression model for complex surveys. By incorporating an adjustment term to the weighted estimating equations, we derived a bias correction based on univariate Bernoulli distributions. This bias correction turns each observation into two: the original response, y_i with the original sampling weight, w_{hij}, times a multiplicative factor, $1+q_{hij}/2$, and a pseudo-response, $1-y_{hij}$ with one minus the original weight times one plus the multiplicative factor, or $q_{hij}/2$. Since both the response and pseduo-response have weights that are guaranteed to be positive, the issue of separation is eliminated. These pseudo-responses and weights are relatively simple to calculate, and this approach leads to an iterative algorithm that is straightforward to implement. Because WEE is the most widely used estimation approach for logistic regression models in complex surveys, the approach to correct for bias described in our manuscript should be useful to applied statisticians.

Although not specifically discussed in this paper, the proposed method can also be used for any model for binary outcomes in complex surveys, including

using those with non-canonical link functions, such as probit or complementary log-log links. Kosmidis (2009) described bias-corrected estimating equations for non-canonical links for binary data with independent observations where the original observations are split into y_i and $1 - y_{hij}$ with weights for each that are a function of a_{hij} in (3.2). Our approach could be an extension of these results seen in Kosmidis and Firth (2011) by incorporating the sampling weights into the appropriate multiplicative factor. Further, this approach can be extended to other generalized linear models using weighted estimating equations for complex surveys. The bias-corrected approach for complex surveys would be similar to that given in Kosmidis and Firth (2009) for other generalized linear models based on specific link functions. This can be done by creating a psuedo-response (a function of the outcome and a_{hij}) to correct for the first-order bias.

Finally, the results of the simulations demonstrate that the proposed method can greatly reduce the finite sample bias of WEE for estimating regression parameters for binary data in the complex survey setting. WEE estimates can be biased due to the issue of separation or quasi-separation, which can occur in large complex surveys when domains are small or subrgoup analyses are performed. The bias-corrected methods perform discernibly better than the standard WEE approach for binary data, suggesting that the bias-corrected method proposed here could be adopted as a first choice in regression analyses of binary outcomes in complex surveys.

Table 3.3: Average relative bias and mean square error of $\hat{\beta}_2$ and empirical coverage probabilties of confidence intervals for each simulation sepcification where the true parameter values are $\beta_2 = \ln(2) \approx 0.69$.

Configuration		Method	Average Relative Bias	Mean Square Error	Empirical Coverage Prob.
$m_i = 5$	$\phi = 0.7$, n = 40	WEE	-75.79	0.045	0.548
		Independence	2.54	0.114	0.947
		Sandwich	1.96	0.107	0.953
		Morel	2.48	0.143	0.989
	$\phi = 0.7$, n = 80	WEE	-75.86	0.065	0.380
		Independence	0.99	0.036	0.95
		Sandwich	0.97	0.05	0.972
		Morel	0.98	0.021	0.954
	$\phi = 0.9$, n = 40	WEE	-85.2	0.322	0.154
		Independence	-7.58	0.157	0.96
		Sandwich	-7.57	0.148	0.949
		Morel	-7.58	0.157	0.982
	$\phi = 0.9$, n = 80	WEE	-84.92	0.331	0.053
		Independence	-2.19	0.089	0.959
		Sandwich	-1.81	0.058	0.99
		Morel	-2.23	0.078	0.952
$m_i = 10$	$\phi = 0.7$, n = 40	WEE	-89	0.725	0.206
		Independence	-3.46	0.095	0.971
		Sandwich	-3.56	0.069	0.983
		Morel	-3.5	0.076	0.979
	$\phi = 0.7$, n = 80	WEE	-89.25	0.764	0.074
		Independence	1.23	0.031	0.948
		Sandwich	1.43	0.033	0.98
		Morel	1.2	0.036	0.966
	$\phi = 0.9$, n = 40	WEE	-90.1	0.642	0.118
		Independence	-3.19	0.12	0.908
		Sandwich	-3.22	0.121	0.935
		Morel	-3.2	0.129	0.962
	$\phi = 0.9$, n = 80	WEE	-89.78	0.628	0.032
		Independence	-2.91	0.043	0.959
		Sandwich	-2.61	0.063	0.967
		Morel	-2.91	0.037	0.982

Based on 2000 replications for each simulation for varying levels of the number of observations in each cluster m_i, levels of ϕ for the bridge distribution, and number of clusters sampled, n.

Table 3.4: Average relative bias and mean square error of $\hat{\beta}_2$ and empirical coverage probabilties of confidence intervals for each simulation sepcification where the true parameter values are $\beta_2 = \ln(4) \approx 1.39$.

Configuration		Method	Average Relative Bias	Mean Square Error	Empirical Coverage Prob.
$m_i = 5$	n = 40, $\phi = 0.7$	WEE	-86.04	2.404	0.192
		Independence	-2.22	0.227	0.908
		Sandwich	-2.04	0.227	0.939
		Morel	-2.24	0.266	0.97
	n = 80, $\phi = 0.7$	WEE	-86.21	2.407	0.042
		Independence	-3.46	0.063	0.96
		Sandwich	-2.75	0.077	0.971
		Morel	-3.45	0.091	0.959
	n = 40, $\phi = 0.9$	WEE	-85.16	1.807	0.108
		Independence	-5.66	0.194	0.928
		Sandwich	-5.9	0.215	0.972
		Morel	-5.59	0.192	0.978
	n = 80, $\phi = 0.9$	WEE	-84.99	1.788	0.031
		Independence	1.05	0.085	0.994
		Sandwich	1.57	0.077	0.994
		Morel	1.01	0.071	0.963
$m_i = 10$	n = 40, $\phi = 0.7$	WEE	-86.55	2.686	0.143
		Independence	-1.54	0.081	0.969
		Sandwich	-1.23	0.106	0.957
		Morel	-1.53	0.075	0.956
	n = 80, $\phi = 0.7$	WEE	-86.43	2.671	0.072
		Independence	1.47	0.054	0.922
		Sandwich	1.94	0.028	0.939
		Morel	1.46	0.042	0.962
	n = 40, $\phi = 0.9$	WEE	-84.95	1.824	0.271
		Independence	0.72	0.058	0.918
		Sandwich	1.03	0.061	0.972
		Morel	0.7	0.061	0.983
	n = 80, $\phi = 0.9$	WEE	-85.21	1.816	0.043
		Independence	-1.52	0.003	0.945
		Sandwich	-1.16	0.046	0.957
		Morel	-1.48	0.026	0.978

Based on 2000 replications for each simulation for varying levels of the number of observations in each cluster m_i, levels of ϕ for the bridge distribution, and number of clusters sampled, n.

Table 3.5: Average relative bias and mean square error of $\hat{\beta}_2$ and empirical coverage probabilties of confidence intervals for each simulation sepcification where the true parameter values are $\beta_2 = \ln(16) \approx 2.77$.

Configuration	Method	Average Relative Bias	Mean Square Error	Empirical Coverage Prob.
$m_i = 5$, n = 40, $\phi = 0.7$	WEE	-92.5	12.53	0.073
	Independence	1.1	0.835	0.979
	Sandwich	11.63	12.088	0.891
	Morel	0.81	0.825	0.909
$m_i = 5$, n = 80, $\phi = 0.7$	WEE	-92.47	12.511	0.032
	Independence	-1.24	0.281	0.958
	Sandwich	-0.66	0.261	0.956
	Morel	-1.2	0.253	0.985
$m_i = 5$, n = 40, $\phi = 0.9$	WEE	-92.67	8.904	0.092
	Independence	4.04	0.762	0.948
	Sandwich	4.31	7.625	0.886
	Morel	3.8	0.747	0.917
$m_i = 5$, n = 80, $\phi = 0.9$	WEE	-92.67	8.914	0.011
	Independence	-1.1	0.238	0.965
	Sandwich	-0.38	0.287	0.985
	Morel	-1.1	0.265	0.993
$m_i = 10$, n = 40, $\phi = 0.7$	WEE	-93.41	12.303	0.231
	Independence	-1.7	0.467	0.943
	Sandwich	-2.94	0.386	0.958
	Morel	-1.84	0.494	0.94
$m_i = 10$, n = 80, $\phi = 0.7$	WEE	-93.52	12.333	0.119
	Independence	-0.92	0.171	0.939
	Sandwich	-0.28	0.159	0.951
	Morel	-0.9	0.168	0.986
$m_i = 10$, n = 40, $\phi = 0.9$	WEE	-91.99	8.156	0.327
	Independence	0.26	0.251	0.955
	Sandwich	0.38	0.264	0.973
	Morel	0.23	0.265	0.942
$m_i = 10$, n = 80, $\phi = 0.9$	WEE	-92.09	8.146	0.013
	Independence	-0.72	0.106	0.947
	Sandwich	-0.19	0.134	0.955
	Morel	-0.71	0.101	0.979

Based on 2000 replications for each simulation for varying levels of the number of observations in each cluster m_i, levels of ϕ for the bridge distribution, and number of clusters sampled, n.

References

Albert, A. and Anderson, J.A. (1984). "On the existence of maximum likelihood estimates in logistic regression models." *Biometrika* **71**, 1–10.

Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–277.

Box, M.J.(1971). "Bias in nonlinear estimation." *Journal of the Royal Statistical Society. Series B (Methodological)*, 171–201.

Bull, Shelley B., Lewinger, J.P., and Lee, S. (2007). "Confidence intervals for multinomial logistic regression in sparse data." *Statistics in Medicine* **26**, 903–918.

Cai, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151–64.

Cai, J., et. al. (2007). Hazard models with varying coefficients for multivariate failure time data. *Annals of Statistics* **35**, 324–354.

Chambers, J. M., Mallows, C. L., and Stuck, B. W. (1976). A method for simulating stable random variables. *Journal of the American Statistical Association* **71**, 340–344.

Cox, D.R. (1975). Partial Likelihood. *Biometrika* **62**, 269–276.

FIRTH, D. (1993). "Bias reduction of maximum likelihood estimates." *Biometrika* **80**, 27–38.

FREEDMAN, LS. (1982). Tables of the number of patients required in clinical trials using the logrank test. *Statistics in Medicine* **1**, 121–129.

HARRINGTON, D. P., FLEMING, T. R.; (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.

HAUCK, W.W. andDONNER, A. (1977). "Wald's test as applied to hypotheses in logit analysis." *Journal of the American Statistical Association* **72**, 851–853.

HCUP, NIS. (2007). "Database Documentation." Healthcare Cost and Utilization Project (HCUP). SID Database Documentation. Rockville, MD: *Agency for Healthcare Research and Quality.*

HEINZE, G. and SCHEMPER, M.(2003). "Comparing the importance of prognostic factors in Cox and logistic regression using SAS." *Computer Methods and Programs in Biomedicine* **71**, 155–163.

HOEL, P.G., PORT, S.C., and STONE, C.J. (1971). *Introduction to Probability Theory.* Boston, MA: *Houghton Mifflin Company.*

HOUGAARD, P. (1986). A class of multivariate failure time distributions. *Biometrika* **73**, 671–678.

HOUGAARD, P. (2000). Analysis of Multivariate Survival Data. *Springer-Verlag.*

HUSTER, W.J., BROOKMEYER, R., & SELF, S.G. (1989). Modelling paired survival data with covariates. *Biometrics* **45**, 145–156.

JUNG, SH. (2007). Sample size calculation for weighted rank tests comparing survival distributions under cluster randomization: a simulation method. *Journal of Biopharmaceutical Statistics* **17**, 839–849.

Jung, SH. and Jeong, JH. (2003). Rank Tests for Clustered Survival Data. *Lifetime Data Analysis* **9**, 21–33.

Kalbfleisch, J. D., Prentice, R. L. (2002). The Statistical Analysis of Failure Time Data. *Wiley*.

Kaplan, E. L.; Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457-481.

Kish, L. (1965). *Survey Sampling*. New York, NY: *Wiley and Sons*.

Klein, J. P., Moeschberger, M.L (2003). Survival Analysis: Techniques for Censored and Truncated Data. *Springer*.

Kosmidis, I. (2009). "On iterative adjustment of responses for the reduction of bias in binary regression models." Technical Report 09–36, *CRiSM working paper series*.

Kosmidis, I. and Firth, D.(2009)."Bias reduction in exponential family non-linear models." *Biometrika***96**, 793–804.

Kosmidis, I. and Firth, D. (2011). "Multinomial logit bias reduction via the Poisson log-linear model." *Biometrika* **98**, 755–759.

Liang, K. Y., Self, S. G., Bandeen-Roche, K. J., Zeger, S. L. (1995). Some recent developments for regression analysis of multivariate failure time data. *Lifetime Data Analysis* **1**, 403–415.

Liang, K. Y., Self, S. G., and Chang, Y. C. (1993). Modelling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society, series B* **55**, 441–453.

Manatunga, A.K. and Chen, S. (2000). Sample Size Estimation for Survival Outcomes in Cluster-Randomized Studies with Small Cluster Sizes. *Biometrics* **56**, 616–621.

MARSHALL, A. W. and OLKIN, I. (1988). .Families of multivariate distributions. *Journal of the American Statistical Association* **84**, 487–493.

MCNEIL, A., FREY, R., and EMBRECHTS, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools. Princeton University Press.*

MOREL, J. G., BOKOSSA, M. C., and NEERCHAL, N. K. (2003). "Small sample correction for the variance of GEE estimators." *Biometrical Journal* **45**, 395–409.

NATARAJAN, S., et. al. (2008). Variance estimation in complex survey sampling for generalized linear models. *Applied Statististician* **65**, 75–87.

PARIKH, A., et.al. (2009). Association Between a DASH-Like Diet and Mortality in Adults With Hypertension: Findings From a Population-Based Follow-Up Study. *American Journal of Hypertension*, **22**, 409–416.

PARZEN, M. et al. (2011). "A generalized linear mixed model for longitudinal binary data with a marginal logit link function." *The Annals of Applied Statistics* **5**, 449.

PETTITT, A. N. (1984). Proportional Odds Models for Survival Data and Estimates Using Ranks. *Journal of the Royal Statistical Society. Series C*, **33**, 169–175.

PETO, R. and PETO, J. (1972). Asymptotically Efficient Rank Invariant Test Procedures. *Journal of the Royal Statistical Society. Series A*, **135**, 185–207.

PRENTICE, R.L. (1978). Linear rank tests with right censored data. *Biometrika*, **65**, 167–179.

PRENTICE, R.L. and HSU, L.(1997). Regression on hazard ratios and cross ratios in multivariate failure time analysis. *Biometrika*, **84**, 349–363.

Rotnitzky, A. and Jewell, N.P. (1990). Hypothesis-Testing of Regression Parameters in Semiparametric Generalized Linear-Models for Cluster Correlated Data. *Biometrika*, **77**, 485–497.

Rubin, D.B. (1997). Estimating Causal Effects from Large Data Sets Using Propensity Scoresa. *Biometrika*, **127**, 757–763.

Schoenfeld, D. (1983). Sample size formula for the proportional-hazards regression model. *Biometrics* **39**, 499–503.

Segal, M.R., Neuhaus, J.M., and James, I.R. (1997). Dependence estimation for marginal models of multivariate survival data. *Lifetime Data Analysis* **3**, 251–268.

Shah, B.V., Barnwell, B.G., and Bieler, G.S.(1996). SUDAAN, Software for the Statistical Analysis of Correlated Data: User's Manual, Release 7.0. *Research Triangle Institute*.

Shih, J.H. and Louis, T.A. (1995). Inferences on the association parameter in copula models for bivariate survival data. *Biometrics* **51**, 1384–1399.

Vittinghoff, E. and McCulloch, C.E. (2007). "Relaxing the rule of ten events per variable in logistic and Cox regression." *American Journal of Epidemiology* **165**, 710–718.

Wang, Z. and Louis, T.A. (2003). "Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function." *Biometrika* **90**, 765–775.

Wei, L.J., Lin, D.Y., and Weissfeld L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of The American StatisticalAssociation*, **84**, 1065–1073.

Xie, J. and Liu, C. (2005). Adjusted KaplanMeier estimator and log-rank test with inverse probability of treatment weighting for survival data. *Statistics in Medicine*, **24**, 3089–3110.

Xie, T. and Waksman, J. (2003). Design and sample size estimation in clinical trials with clustered survival times as the primary endpoint. *Statistics in Medicine* **22**, 2835–2846.

Yu, H. et al. (2012). "Comparative analysis of outcomes and costs following open radical cystectomy versus robot-assisted laparoscopic radical cystectomy: results from the US Nationwide Inpatient Sample." *European Urology* **61**, 1239–1244.